ÉVALUATION

Fondements, controverses, perspectives

Sous la direction de

Thomas Delahais,
Agathe Devaux-Spatarakis,
Anne Revillard et Valéry Ridde

ÉVALUATION

ÉVALUATION

Fondements, controverses, perspectives

Sous la direction de Thomas Delahais,
Agathe Devaux-Spatarakis, Anne Revillard et
Valéry Ridde

QUÉBEC : ÉDITIONS SCIENCE ET BIEN COMMUN

Titre : Évaluation. Fondements, controverses, perspectives

Sous la direction de Thomas Delahais, Agathe Devaux-Spatarakis, Anne Revillard et Valéry Ridde

Design de la couverture : Kate McDonnell

Édition et révision linguistique : Alexandre Prince, Érika Nimis

ISBN pour l'impression : 978-2-925128-14-4

ISBN pour le ePub : 978-2-925128-16-8

ISBN pour le PDF : 978-2-925128-15-1

Dépôt légal – Bibliothèque et Archives nationales du Québec 2021

Dépôt légal – Bibliothèque et Archive nationale Canada

Éditions science et bien commun
http://editionscienceetbiencommun.org
3-855 avenue Moncton
Québec (Québec) G1S 2Y4

Diffusion : info@editionscienceetbiencommun.org

Table des matières

Introduction générale

THOMAS DELAHAIS, AGATHE DEVAUX-SPATARAKIS, ANNE REVILLARD ET VALÉRY RIDDE

Au moment où nous avons lancé l'idée d'élaborer et de partager cette somme de textes fondamentaux en évaluation, le monde n'était pas encore aux prises avec une nouvelle pandémie ravageuse. La crise sanitaire mondiale liée à l'épidémie de SARS-CoV-2 a, entre-temps, révélé avec une acuité particulière la pertinence des questionnements analysés dans le champ de l'évaluation des interventions, mais aussi la diffusion encore trop confidentielle des acquis de cette discipline. Comment les actions initiées et réalisées par les gouvernements, les organisations et les populations ont-elles été en mesure d'endiguer la pandémie, d'en réduire les effets et leur répartition inégale entre différents groupes? Comment ces interventions ont-elles pris en compte l'état des connaissances scientifiques pour favoriser leur mise en œuvre, leur efficacité? Dans quelle mesure les responsables politiques ont-ils et elles intégré les apports des travaux d'évaluation dans leurs décisions?

Comme on le voit, si la pandémie est révélatrice des nombreux maux affectant le monde depuis des siècles, elle permet une nouvelle fois de poser la question de la place de l'évaluation dans nos sociétés, nos décisions, nos connaissances, nos enseignements et nos relations sociales. En effet, les multiples débats et controverses dans le contexte de la pandémie sur l'efficacité de telle intervention (par exemple, le couvre-feu) ou de tel traitement (l'hydroxychloroquine), ou encore sur l'impossible évaluation des actions entreprises (Paul et Ridde, 2020) du fait de leur complexité (Pawson, Manzano et Wong, 2020) montrent que non seulement la culture scientifique des responsables politiques reste faible, mais aussi que leurs connaissances en évaluation le sont tout autant.

Le contexte sanitaire a ainsi mis un coup de projecteur sur le constat à l'origine de ce recueil, à savoir le déficit de connaissances fondamentales et historiques concernant le domaine de l'évaluation. Nous utiliserons indifféremment les termes « évaluation de programmes », « évaluation des interventions » ou « évaluation des politiques publiques », pour désigner le champ de pratique professionnelle et de recherche qui s'est développé depuis une cinquantaine d'années autour de la détermination de la valeur d'un certain nombre d'interventions publiques, ou visant des objectifs d'intérêt général.

L'évaluation résiste depuis longtemps à une définition simple et unanime, qui refléterait à la fois son histoire, et la diversité des acteurs, des usages et des contextes dans lesquels elle s'inscrit. Notre objet n'est pas ici d'en formuler la définition ultime. Toutefois, nous observons que la plupart des définitions comprennent trois éléments indispensables (Demarteau, 2002) : l'élaboration d'un jugement sur la valeur, le recours à des méthodes spécifiques dans une démarche d'investigation systématique et la volonté de favoriser l'utilité des travaux. Nous y ajouterions un quatrième : le souci de l'intérêt général ou du bien commun – ou, plus explicitement, de justice sociale (Mertens et Wilson, 2012). Ces « ingrédients » constituent la base d'une infinité de recettes et, partant, de débats sur l'évaluation en théorie et en pratique, lesquels trouveront une abondante illustration dans cet ouvrage.

Ainsi, la définition de la Société américaine d'évaluation (AEA) se place fortement sous l'égide de la valeur, dans la suite de Michael Scriven : « l'évaluation est le processus par lequel on détermine le mérite, l'intérêt (*worth*) et l'importance (*significance*) des choses » (Scriven, 1991 : 1). La notion de « mérite » renvoie aux qualités intrinsèques d'une intervention, « l'intérêt » à son apport dans un contexte précis, et « l'importance » à un jugement plus global sur l'intervention combinant mérite et intérêt – le tout constituant sa valeur (*value*). La particularité de l'évaluation se trouve dans le processus spécifique mis en place pour définir les critères au regard desquels la valeur est déterminée – processus ouvert

et observable, impliquant de façon transparente des acteurs donnés, aboutissant à des critères explicites, et à une stratégie pour y répondre (Barbier, 1990 : 34).

D'autres définitions font de prime abord référence aux méthodes. Pour Peter Rossi *et al.*, « l'évaluation de programme est le recours à des procédures issues des sciences sociales pour investiguer systématiquement l'efficacité des interventions sociales » (Rossi *et al.*, 2004 : 4). La définition de la Société suisse d'évaluation, la SEVAL, s'inscrit elle aussi dans cette tradition, en définissant l'évaluation comme « une analyse et une appréciation systématique et transparente de la conception, de la mise en œuvre et/ou des effets d'un objet d'évaluation » (2016).

L'insistance sur l'utilité se retrouve dans la définition donnée par la Société française d'évaluation (2006) : « L'évaluation vise à produire des connaissances sur les actions publiques, notamment quant à leurs effets, dans le double but de permettre aux citoyens d'en apprécier la valeur et d'aider les décideurs à en améliorer la pertinence, l'efficacité, l'efficience, la cohérence et les impacts ». Si cette définition constitue un compromis entre une vision démocratique et une vision comptable de l'évaluation, notons que l'insistance sur la notion d'utilité pour certaines parties prenantes (qu'il s'agisse des décideurs et des décideuses, des opérateurs et opératrices, des intervenant-e-s de première ligne ou des citoyen-ne-s en général) est un élément fondateur de nombreuses approches évaluatives.

Ces différentes définitions ne s'opposent pas et constituent autant de constats valables concernant la théorie et la pratique évaluative. Pour en revenir à notre métaphore culinaire, c'est le poids respectif de chaque ingrédient, l'ordre et les modalités selon lesquels ils sont incorporés, qui rendent compte de la richesse du champ investigué.

Un accès difficile aux écrits anglophones sur l'évaluation

En 1995, Jacques Toulemonde s'intéressait à l'émergence d'une profession évaluative en Europe et identifiait quatre catégories de producteurs et productrices d'évaluation (Toulemonde, 1995) : les professionnel-le-s, les spécialistes (qui connaissent l'évaluation mais n'en font pas leur activité principale), les artisan-e-s (qui connaissent les techniques et ont une forte expertise d'usage, mais en ignorent les fondements théoriques) et les amateurs et amatrices. Il pensait alors assister à un glissement des amateurs/-trices vers les artisan-e-s, et des artisan-e-s vers les spécialistes. Vingt-cinq ans après, il n'est pas certain que, dans l'Europe francophone tout du moins, cette évolution ait porté ses fruits. Comme le constatait déjà Lee J. Cronbach en 1980, les personnes qui évaluent, qu'elles soient chercheur-e-s, consultant-e-s, ou fonctionnaires, continuent souvent à importer dans l'évaluation leurs propres cadres de pensée, qu'ils soient théoriques, conceptuels ou pratiques, et à considérer l'évaluation comme une activité technique et méthodologique (mener des entretiens, des enquêtes, écrire un rapport etc.) sans tenir compte de l'histoire ou des dernières réflexions concernant ce champ de pratiques.

De nombreux facteurs contribuent à une telle méconnaissance, qui trouve ses racines dans l'enseignement secondaire et supérieur, dans la place de la science et de l'évaluation dans les discours et les pratiques, dans les priorités budgétaires et parlementaires, dans la professionnalisation des pratiques, etc. En ce qui nous concerne, si l'histoire de l'étude de l'action publique est relativement connue par les francophones (Laborier et Trom, 2003), celle de l'évaluation l'est certainement moins, à l'exception des ouvrages classiques d'introduction au sujet (Beaudry et Gauthier, 1992; Dagenais et Ridde, 2009; Perret, 2008) qui, faute de place, ont trop souvent fait l'impasse sur les écrits plus anciens. La faible accessibilité, pour un public francophone, des travaux internationaux en évaluation, notamment les plus anciens, est un enjeu

majeur, ces travaux se trouvant à l'origine des réflexions et des pratiques actuelles. De nombreux freins existent quant à l'accès à ces écrits. Des freins linguistiques évidemment, tant il est vrai que le monde francophone éprouve encore des difficultés à lire la production anglophone. Mais aussi des freins matériels relevant de l'accession inégale du plus grand nombre à la connaissance (Piron *et al.*, 2016), puisque la majeure partie de ces écrits sont encore diffusés par des éditeurs commerciaux qui en restreignent l'accès, imposant aux auteurs et autrices de céder leur droit de diffusion gratuite de leurs productions. De ce fait, on s'étonne parfois d'assister à des débats francophones dans le domaine de l'évaluation sur des sujets qui ont été abordés depuis longtemps dans le monde anglophone[1]. Un des objectifs de cette compilation est donc de favoriser une meilleure appropriation par les praticien-ne-s et théoricien-ne-s francophones en évaluation, des acquis des écrits anglophones à partir de quelques branches de « l'arbre de l'évaluation », et de ses trois grandes ramifications autour des méthodes, des valeurs et des usages de l'évaluation (Christie et Alkin, 2012).

Ajoutons que les efforts pour importer dans la pratique de l'évaluation des cadres théoriques issus du monde de la recherche n'ont pas toujours été soutenus. Ce déficit concernent également certaines équipes de recherche en santé mondiale (Ridde, Pérez et Robert, 2020) ou spécialisées dans l'étude des politiques publiques (Jones, Gautier et Ridde, 2021). Si les guides et modes d'emploi se multiplient, et sont maintenant de bonne qualité, ils tendent à présenter l'évaluation comme une succession d'étapes, sans la situer dans un ensemble plus vaste, et sans montrer la concomitance et l'imbrication de ses sous-processus (Dagenais et Ridde, 2009). Outre la barrière de la langue, les écrits en évaluation ne sont pas nécessairement d'un accès facile, du fait de leur nombre, et d'une tendance à l'inflation conceptuelle. Comment se repérer dans ce foisonnement? Et comment appréhender cet objet pour des

1. L'inverse est évidemment aussi parfois valable, et nous n'avons ni préjugé ni complexe vis-à-vis de nos collègues qui n'écrivent qu'en anglais!

francophones peu à l'aise dans la lecture de l'anglais? S'il est vrai que les outils de traduction automatique sont de plus en plus performants, il n'en reste pas moins que ceux-ci ne permettent pas de rendre compte de l'« épaisseur conceptuelle » des cadres et des théories (Berthoud, 2018). Ainsi, l'accès aux cadres théoriques est marqué par une triple barrière : l'accès aux écrits scientifiques, la langue et le champ lexical de la discipline. Notre recueil de traductions d'une série de textes anglophones ayant marqué l'histoire du champ de l'évaluation et posant les jalons de ses perspectives d'évolution vise ainsi à combler ce manque.

Un besoin de réflexivité et d'historicité

Au-delà du seul travail de traduction, il s'agit donc, dans cet ouvrage, d'accompagner les lecteurs et les lectrices dans l'exploration du champ de l'évaluation, en rendant compte des aspects des principaux débats qui le traversent. Nous considérons que pour les professionnel-le-s de l'évaluation, replacer celle-ci dans son cadre théorique et conceptuel permet de redonner du sens à la pratique, et d'échapper au risque de routine évaluative dans lequel peuvent être pris-e-s les praticien-ne-s, ou les organisations qui réclament des évaluations. Notre visée est d'enrichir les pratiques, et de permettre à l'évaluation de modifier en profondeur les points de vue sur les politiques publiques. Cela n'est possible qu'en combinant robustesse méthodologique et cadres évaluatifs de qualité. S'adressant également à un public académique, cet ouvrage cherche à mettre en lumière les apports des « théories issues de la pratique »[2] de l'évaluation, sur des questions que la recherche aborde parfois moins frontalement, telles que : quelle est l'utilité des nouvelles connaissances obtenues? Ou encore : à l'aune de quelles valeurs sont-elles produites?

2. Pour reprendre l'expression de William Shadish, Thomas Cook et Laura Leviton dans le titre de leur ouvrage *Foundations of Program Evaluation: Theories of Practice*, paru chez Sage (Londres) en 1991.

Notre ouvrage s'adresse bien sûr également aux commanditaires de l'évaluation, mais aussi et surtout à toutes celles et tous ceux qui sont concerné-e-s, affecté-e-s, subissent ou profitent des interventions publiques ou d'intérêt – c'est-à-dire chacun-e d'entre nous : il s'agit en effet de replacer l'évaluation dans le débat démocratique, afin de ne pas reproduire la défaillance constatée lors de la lutte contre la pandémie. Démocratiser l'évaluation, c'est proposer aux citoyen-ne-s des ressources intellectuelles pour réfléchir aux enjeux des interventions publiques, et disposer de diagnostics pertinents sur la mise en œuvre et les effets de celles-ci. Autrement dit, c'est ouvrir la discussion sur les valeurs, tout en appuyant les opinions avec des faits.

C'est donc à ce travail de sélection, de traduction, de synthèse et de mise en perspective que nous nous sommes attelé-e-s dans cet ouvrage, que nous proposons en accès libre. Nous avons sélectionné les grands thèmes et extraits de textes qui vont suivre, avec pour objectif de rendre compte de la longue histoire des débats évaluatifs. Étant donné la richesse du champ, la sélection n'a pas été facile. Nous avons notamment pris le parti de nous concentrer sur les enjeux généraux de la *démarche* d'évaluation (les valeurs qu'elle engage, son utilité, ses acteurs et actrices, ses liens avec la recherche), sans entrer dans le détail technique des *méthodes* d'évaluation[3]. Précisons ici que les textes traduits ne défendent pas nécessairement des positions que nous partageons; ils ont vocation à rendre compte de la diversité et de la richesse des débats qui traversent le champ de l'évaluation, pour permettre aux lectrices et aux lecteurs d'en saisir les enjeux et de définir le cas échéant leur propre position. Loin de constituer une liste arrêtée de références incontournables, ces textes doivent être pris comme autant d'éclairages que nous proposons sur les enjeux de la démarche d'évaluation de programme. Ajoutons que nous

3. Des approches de type expérimental (expérimentations avec assignation aléatoire, aussi appelée essais randomisés contrôlés ou ERC) aux démarches plus qualitatives d'analyse de la théorie du programme (aussi désignée sous les noms de théorie de l'intervention ou théorie du changement), le champ de l'intervention a été un lieu important d'innovation méthodologique.

avons veillé à limiter la longueur des textes traduits, afin d'en faciliter l'accès pour les praticien-ne-s, mais aussi pour les chercheuses et chercheurs déjà confronté-e-s à l'abondance des ressources relevant de leur discipline d'appartenance. Ainsi, nous avons souhaité proposer une diversité d'approches et de réflexions, tout en donnant une dimension historique à notre démarche.

Ces textes ont été traduits, pour permettre aux lecteurs et lectrices francophones d'apprécier les débats qui agitent la communauté scientifique évaluative à l'échelle internationale. Le montant exorbitant des droits de traduction demandés par certains éditeurs (allant jusqu'à 29 750€ pour un article de 16 pages!) nous a conduit-e-s à abandonner certains textes initialement sélectionnés, ce qui montre, s'il en était besoin, la pertinence du combat actuel pour un libre accès aux publications scientifiques[4].

Enfin, pour mettre en perspective les quelque cinquante années de réflexion évaluative retracées par ces textes, des chercheurs, chercheuses et praticien-ne-s francophones contemporain-e-s (Manuela De Allegri, Tom Archibald, Yves Gingras, Marthe Hurteau, Nathalie Mons) ont généreusement accepté de discuter chacune des cinq parties. Ces discussions replacent les textes dans les débats actuels et introduisent de nouvelles perspectives, distinctes des nôtres. Nous souhaitons ainsi contribuer à ancrer la traduction comme processus de réflexion pour les évaluateurs/-trices nouveaux et nouvelles, souhaitant entrer dans ce champ; pour les autres, nous aimerions promouvoir durablement la réflexivité issue les des pratiques évaluatives comme processus essentiel à toute pratique professionnelle (Alexander *et al.*, 2020).

4. Le Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP) a dépensé un total de 5 560€ en droits de traduction versés aux différents éditeurs concernés.

Ainsi, grâce à cet ouvrage, nous espérons ouvrir des chemins d'exploration aux francophones pratiquant ou étudiant l'évaluation. Pour les personnes ayant déjà une expérience de l'évaluation, la lecture des textes choisis permettra de mettre en perspective les questionnements rencontrés dans de leur pratique, voire d'en susciter de nouveaux. Notre objectif est également de leur permettre de mieux se positionner parmi la diversité des points de vue tout en tenant compte de leur parcours sans avoir à relire toute la généalogie de l'évaluation, de ses fondements (Alkin, 2004) à ses derniers développements (Lemire, Peck et Porowski, 2020). Nous espérons ouvrir aux novices en évaluation le champ des possibles tant dans la pratique que dans la réflexion évaluative, en y incluant les conséquences des choix qu'ils et elles pourront être amené-e-s à faire dans l'exercice de ce qui est un métier à part entière. Libre ensuite à elles et eux d'élargir ou de prolonger les chemins qui contribuent à enrichir leur réflexion professionnelle En effet, bien loin d'une somme finie, ce recueil doit être considéré comme une pierre à l'édifice de la réflexion collective du monde francophone de l'évaluation.

Des textes essentiels concernant cinq domaines en évaluation

Ce recueil est divisé en cinq parties représentant cinq domaines essentiels à la compréhension du champ de la pratique de l'évaluation. Celles-ci peuvent être abordées indépendamment les unes des autres, et dans l'ordre souhaité par le lecteur ou la lectrice en fonction de ses intérêts personnels.

Partie 1/ À quoi sert l'évaluation?

Dans cette partie (à laquelle il sera fait référence dans le reste du texte sous le nom de « Partie Utilité »), nous présentons les travaux canoniques s'intéressant à l'utilité de l'évaluation, ou proposant des pratiques qui visent à la promouvoir. Ces réflexions sont centrales pour la communauté évaluative puisqu'elles interrogent la raison d'être de cette activité ainsi que sa place dans l'action publique. Considérer l'utilité de l'évaluation uniquement à l'aune de la prise en compte des rapports d'évaluation par les responsables politiques a graduellement fait place à une caractérisation de la diversité des usages. Cette diversité pose alors la question du degré d'influence que l'évaluation peut être en mesure d'exercer sur l'action publique, et des mécanismes pouvant la favoriser. Enfin, ces réflexions peuvent aussi fournir l'occasion de proposer des pratiques évaluatives mettant explicitement le développement des principes démocratiques et l'amélioration du bien-être de tou-te-s les citoyen-ne-s au cœur de leur action.

Partie 2/ Qui évalue et comment?

Qui évalue? Quel est le rôle de celles et ceux qui « font les évaluations » dans la pratique évaluative? Dans cette partie (« Partie Évaluatrice[5] ») nous nous intéressons à ces différentes questions en recontextualisant la transformation de l'évaluation. D'une activité parmi d'autres réalisée principalement par des équipes issues du monde académique, celle-ci est devenue une pratique, et même un métier en tant que tel, ouverts à un nombre grandissant d'acteurs et d'actrices : fonctionnaires, consultant-e-s, personnels associatifs. Si les premier-e-s évaluateurs/-trices se voyaient d'abord et avant tout comme les garant-e-s d'une objectivité

5. Nous optons ici pour un féminin générique afin d'alléger l'écriture.

scientifique permettant d'adosser la prise de décision à une vérité, très vite s'est posée la question de leur rôle véritable pour une évaluation utile, à même de prendre en compte le système de valeurs des différentes parties prenantes, et suffisamment robuste en dépit des contraintes (de données, de budget et de temps) souvent très fortes. Cette partie retrace donc les débats qui ont agité la communauté évaluative sur des questions fondamentales : que signifie bien faire son métier? Comment prendre en compte tous les points de vue? L'évaluateur ou l'évaluatrice doit-il ou elle prendre un rôle actif dans la défense du bien commun?

Partie 3/ Évaluer : en fonction de quelles valeurs?

Évaluer les interventions, c'est en déterminer la valeur. Cette partie (« Partie Valeurs ») revient sur ce qui est sans doute un des éléments les plus structurants de la pratique évaluative, à savoir la logique évaluative, c'est-à-dire le processus structuré par lequel des faits collectés dans le cadre de l'évaluation sont jugés à l'aune de « ce qui compte » dans le contexte de l'évaluation. Faut-il procéder à un jugement ou bien doit-on se contenter d'énoncer des faits? Qui est légitime pour porter un jugement? Quelles sont les valeurs au regard desquelles établir ce jugement? Dans cette partie, nous revenons notamment sur le glissement progressif d'une approche selon laquelle il revient à l'équipe d'évaluation de définir les critères de jugement ainsi que les niveaux de performance à atteindre à une démarche dans laquelle la définition des critères et la formulation des jugements deviennent partie intégrante d'un travail participatif incluant les différentes parties prenantes. Nous terminons en dressant un panorama des approches transformationnelles, lesquelles prennent parti pour les valeurs de celles et ceux qui sont dominé-e-s ou subissent les politiques publiques. Ainsi, les évaluations dites féministes, attentives aux différences culturelles, ou autochtones vont plus loin dans le processus visant à changer notre regard sur les interventions.

Partie 4/ L'évaluation est-elle une science?

Dans cette partie (« Partie Science ») nous nous interrogeons sur les liens entre évaluation et recherche scientifique : en quoi l'évaluation relève-t-elle d'une science, et quels sont, le cas échéant, ses apports au champ scientifique? L'évaluation a été historiquement définie comme une pratique de sciences sociales appliquées, mobilisant ses méthodes au service d'une analyse des enjeux et des conséquences des politiques publiques. De ce fait, elle relève d'abord de la science par ses méthodes. Pour autant, son institutionnalisation dans le champ universitaire reste limitée, notamment du fait de son caractère interdisciplinaire et largement appliqué. Or, l'évaluation ne se contente pas d'emprunter à des méthodes scientifiques : elle les enrichit en retour, et approfondit les questionnements concernant les valeurs et les critères de jugement, mais aussi l'utilité des savoirs produits, une notion qui reste trop souvent implicite dans la recherche scientifique. Ce chapitre retrace les enjeux épistémologiques et institutionnels de l'articulation entre évaluation et science, mais aussi les enjeux symboliques, particulièrement saillants dans un contexte où la démarche scientifique elle-même fait l'objet de remises en cause politiques virulentes.

Partie 5/ La pluralité des approches paradigmatiques

La question des paradigmes, centrale dans le domaine de la science, se retrouve en évaluation. Ce champ de pratique fait appel à des méthodes et des raisonnements scientifiques, surtout lorsqu'il est question d'évaluer l'efficacité des interventions, et donc d'effectuer une analyse de la causalité. Ainsi, la manière d'appréhender et de comprendre le monde vient-elle obligatoirement influencer la capacité de l'évaluatrice ou de l'évaluateur à affirmer qu'une intervention a été efficace. Il s'agit d'un sujet complexe, à propos duquel les débats sont aussi nombreux et anciens que

le sont les points de vue. Les textes de cette partie (« Partie Paradigmes ») visent à en montrer la complexité avec une première évocation des approches expérimentales en évaluation, et des types de causalité qu'il est possible de dégager. Ensuite, d'autres textes sont présentés qui permettent de comprendre l'état des controverses sur des enjeux, contemporains en France mais abordés depuis longtemps dans la littérature – citons par exemple les essais contrôlés randomisés ou ECR. Pourtant, il existe une myriade d'approches de la causalité, et les données qualitatives peuvent être d'une grande utilité dans cette perspective. Enfin, il nous a semblé essentiel de donner accès au débat très intense entre celles et ceux qui suggèrent que les différentes approches de la causalité sont irréconciliables, et d'autres défendant un croisement de l'évaluation « expérimentale » avec une perspective sociologique du réalisme critique.

Notre recueil a donc pour ambition d'offrir une lecture historique et diversifiée des fondements, des controverses et des perspectives en évaluation des interventions. Nous croyons à l'importance d'inscrire nos pratiques évaluatives dans l'histoire de ce champ, dans des théories et des cadres conceptuels pertinents, sans s'y enfermer, ni y recourir par réflexe, habitude, opportunisme, lassitude ou obligation. Nous pensons aussi qu'il est essentiel que ces réflexions théoriques et conceptuelles soient nourries par les praticiens et praticiennes de l'évaluation -étudiant-e-s, expert-e-s, consultant-e-s, universitaires, responsables de programmes ou intervenant-e-s de première ligne. Nous espérons que la lecture approfondie et critique de ces textes permettra aux membres de cette vaste, diverse et multicolore communauté francophone de l'évaluation, de s'engager dans un débat à double sens entre pratiques et théories, fondé sur la connaissance des réflexions antérieures que nous avons le plaisir de proposer pour la première fois en français.

Bibliographie

Alexander, Stephanie A., Catherine M. Jones, Marie-Claude Tremblay, Nicole Beaudet, Morten Hulvej Rod et Michael T. Wright. 2020. « Reflexivity in Health Promotion : A Typology for Training ». *Health Promotion Practice* 21(4):499-509. doi : https://doi.org/10.1177/1524839920912407.

Alkin, Marvin C. 2004. *Evaluation Roots. Tracing Theorist's Views and Influences.* 6e éd. Thousand Oaks, Calif : Sage Publications.

Barbier, Jean-Marie. 1990. *L'évaluation en Formation*. Paris : PUF.

Beaudry, Jean, et Benoît Gauthier. 1992. « L'évaluation de programme ». in *Recherche sociale : de la problématique à la collecte de données*, édité par B. Gauthier. Sillery (Québec) : Presses de l'Université du Québec, p.425-52.

Berthoud, Anne-Claude. 2018. « Des pratiques scientifiques plurilingues pour la qualité de la connaissance ». *Repères DoRiF* 17(7).

Christie, Christina A., et Marvin C. Alkin. 2012. « An evaluation theory tree ». in *Evaluation Roots*, édité par C. A. Christie et M. C. Alkin. London: Sage Publications.

Cronbach, Lee J., Sueann R. Ambron, Sanford M. Dornbusch, Robert D. Hess, Robert C. Hornik, Denis Charles Phillips, Decker F. Walker et Stephen S. Weiner. 1980. *Toward Reform of Program Evaluation : Aims, Methods, and Institutional Arrangements*. 1re éd. San Francisco : Jossey-Bass.

Dagenais, Christian, et Valéry Ridde. 2009. *Approaches et pratiques en évaluation de programme*. Montréal : Presses de l'Université de Montréal.

Demarteau, Michel. 2002. « A Theoretical Framework and Grid for Analysis of Programme-evaluation Practices ». *Evaluation* 8(4):454 473. doi : https://doi.org/10.1177/13563890260620649.

Jones, Catherine, Lara Gautier et Valéry Ridde. 2021. « A scoping review of theories and conceptual frameworks used to analyse health financing policy processes in sub-Saharan Africa ». *Health Policy and Planning* 36(7):1197-1214. doi : https://doi.org/10.1093/heapol/czaa173.

Laborier, Pascale, et Danny Trom. 2003. *Historicités de l'action publique*. Paris : Presses Universitaires de France.

Lemire, Sebastian, Laura R. Peck, et Allan Porowski. 2020. « The growth of the evaluation tree in the policy analysis forest: recent developments in evaluation ». *Policy Studies Journal* 48(S1) : S47-70. doi : https://doi.org/10.1111/psj.12387.

Mertens, Donna M., et Amy T. Wilson. 2012. *Program Evaluation Theory and Practice*: A *Comprehensive Guide*. New York : Guilford Press.

Paul, Elisabeth, et Valéry Ridde. 2020. « Evaluer les effets des différentes mesures de lutte contre le Covid-19, mission impossible? » *The Conversation*.

Pawson, Ray, Ana Manzano, et Geoff Wong. 2020. « The Coronavirus response : known knowns, known unknowns, unknown unknowns ».

Perret, Bernard. 2008. *L'évaluation des politiques publiques*. Paris : La Découverte.

Piron, Florence, Samuel Régulus, Marie Sophie Dibounje Madiba, Thomas Hervé Mboa Nkoudou, Dany Rondeau, Marie-Claude Bernard, Jean Jacques Demba, *et al.* 2016. *Justice cognitive, libre accès et savoirs locaux*. Québec : Éditions science et bien commun.

Ridde, Valéry, Dennis Pérez, et Emilie Robert. 2020. « Using implementation science theories and frameworks in global health ». *BMJ Global Health* 5(4). doi : https://doi.org/10.1136/bmjgh-2019-002269.

Rossi, Peter H., Mark W. Lipsey, et Howard E. Freeman. 2004. *Evaluation : A systematic approach*. 7e éd. Thousand Oaks: Sage Publications.

Scriven, Michael. 1991. *Evaluation Thesaurus*. 4e éd. Thousand Oaks : Sage Publications.

Shadish, William R., Thomas D. Cook et Laura C. Leviton. 1991. *Foundations of Program Evaluation: Theories of Practice*. Thousand Oaks: Sage Publications.

Société française d'évaluation (SFE). 2006. « Charte de l'évaluation ».

Société suisse d'évaluation (SEVAL). 2016. « Standards d'évaluation de la Société suisse d'évaluation (Standards SEVAL) ».

Toulemonde, Jacques. 1995. « The emergence of an Evaluation Profession in European Countries : Is there a provision of professionals? » *Knowledge and Policy*.

I. À QUOI SERT L'ÉVALUATION?

Introduction : à quoi sert l'évaluation?

AGATHE DEVAUX-SPATARAKIS, THOMAS DELAHAIS, ANNE REVILLARD ET VALÉRY RIDDE

Après un demi-siècle de financement d'évaluations à travers le monde, quel bilan dresse-t-on de leurs impacts sur l'amélioration de l'action publique? Ce questionnement autour de l'utilité de l'évaluation est central, d'une part parce que cette dernière est principalement financée par des fonds publics, et donc soumise à des obligations de redevabilité; d'autre part parce qu'elle est conduite dans le cadre d'une commande visant à améliorer l'action publique. Ainsi, déterminer si l'évaluation a une influence sur la conduite de l'action publique revient à lui appliquer sa propre démarche, et à se demander dans quelle mesure elle atteint ses objectifs.

Force est de constater qu'à ce jour, l'évaluation n'a pas été déterminante pour la transformation de l'action publique (Spenlehauer, 2016; Weiss, 1988). Le personnel politique n'y fait que rarement référence, et elle reste quasiment absente du débat parlementaire et citoyen. Néanmoins, celle-ci continue d'être financée et promue que ce soit au niveau national ou par les organismes internationaux, notamment par l'OCDE, les Nations Unies ou la Banque Mondiale.

Mais alors, que vise l'évaluation? Faut-il chercher des impacts de l'évaluation ailleurs que dans la prise de décision publique? Quelle place peut-elle revendiquer dans la conduite de l'action publique?

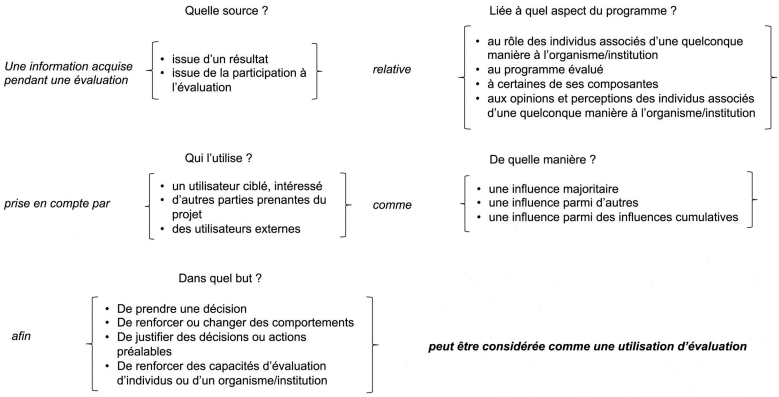
Des décennies d'analyse et de débats sur l'utilisation des évaluations nous permettent d'apporter un éclairage sur ces questionnements.

Fournir une réponse tranchée et rigoureuse sur l'utilité ou l'utilisation de l'évaluation est un exercice périlleux. Cela requiert de définir en amont les objectifs de l'évaluation, et d'en préciser les critères de jugement. Or,

ceux-ci sont légion. Dès ses prémices, l'évaluation a été présentée par ses promoteurs et promotrices comme un instrument au service de la transformation de l'intervention publique, porteuse soit d'une approche expérimentale (Campbell, 1969), soit d'une logique de rationalisation et d'économie des dépenses publiques (dans le cadre du mouvement de la rationalisation des choix budgétaires), voire d'une démarche favorisant la participation citoyenne au contrôle et à la fabrique des politiques publiques (Viveret, 1989). Dès lors, comment proposer une définition univoque de l'utilisation de l'évaluation?

Après une revue extensive des écrits scientifiques à ce sujet, Marvin Alkin et Jean King (Alkin et King, 2017) proposent une définition de l'utilisation de l'évaluation par une phrase composite (notre traduction).

Figure 1 : Définition de l'utilisation de l'évaluation (Alkin et King, 2017)



Ainsi, selon cette définition, « une information acquise pendant une évaluation, [issue d'un résultat], relative [au programme évalué] prise en compte par [un utilisateur ou une utilisatrice ciblé-e, intéressé-e] comme [une influence majoritaire] afin [de prendre une décision] peut être considérée comme une utilisation d'évaluation ». Mais aussi « une information acquise pendant une évaluation, [issue de la participation à

l'évaluation], relative [au rôle des individus associé-e-s d'une quelconque manière à l'organisme/institution] prise en compte par [des utilisateurs et utilisatrices externes] comme [une influence parmi d'autres] afin [de justifier des décisions ou actions préalables] peut de la même manière être considérée comme une utilisation d'évaluation ».

En outre, il s'agit de considérer que chaque possibilité proposée dans cette définition recouvre à elle seule une diversité de contextes démultipliant les types d'utilisation pouvant être observés (Patton, 2020). Dès lors, cette diversité de configurations a entraîné la nécessité de caractériser la multiplicité des usages, et remet en question la possibilité même d'élaborer une grille d'indicateurs standards permettant de poser un diagnostic sur l'utilisation d'une d'évaluation.

À cette multiplicité d'utilisations et d'usages de l'évaluation s'ajoutent les « mésusages » (*misuse*) de l'évaluation. C'est-à-dire l'utilisation intentionnelle de l'évaluation ou de ses résultats à mauvais escient. Cette notion de « mésusages » s'appuie sur une appréciation normative de cet usage, selon la déontologie guidant les pratiques des équipes d'évaluation. L'exemple le plus courant est la sélection des résultats positifs partiels de l'évaluation, et la mise à l'écart de résultats défavorables afin de servir les intérêts politiques (Rajkotia, 2018).

On trouve enfin bien sûr l'absence d'utilisation, laquelle peut être intentionnelle ou pas, suivant que les utilisateurs et utilisatrices ciblé-e-s n'ont pas connaissance de l'évaluation ou que celle-ci n'a pas apporté de démonstration étayée et convaincante (Cousins, 2004).

Les rôles qui incombent à l'équipe d'évaluation dans la promotion de son travail sont multiples, et s'organisent autour de ces débats. Pour certain-e-s, concentrant leurs efforts sur l'utilisation des résultats de l'évaluation, l'équipe d'évaluation peut agir en conduisant une démonstration la plus rigoureuse possible de la causalité, afin de convaincre les décideurs et décideuses (Duflo, 2005). Leur travail se concentre alors sur la scientificité de la méthode. Pour d'autres, visant une multiplicité d'usages

et d'influences, l'équipe d'évaluation peut renforcer l'utilité de ses travaux en étudiant et en adaptant son approche et ses méthodes au contexte de la demande et aux besoins de ses commanditaires (Patton, 1988). Ils s'intéressent donc en priorité à l'analyse du contexte, des enjeux de l'évaluation, et des motivations de la prise de décision.

Fondements : de l'utilisation des résultats aux usages de l'évaluation

Cette problématique a été abordée dans un premier temps à partir du diagnostic à établir sur l'utilité des évaluations. En effet, la proximité de l'évaluation avec la science (cf. partie Science) a d'abord conduit les observateurs et les observatrices à étudier son utilisation en fonction des mêmes critères que ceux des travaux scientifiques (Weiss, 1972). L'emploi de ce prisme d'analyse visait principalement l'utilisation des « livrables de l'évaluation », c'est-à-dire les rapports et les connaissances produites par l'évaluation pour éclairer la décision publique. La prise en compte de ces critères a conclu à la faible utilité de l'évaluation.

Peu à peu, cette approche a fait place à une conception élargie des utilisations possibles de l'évaluation. Des recherches ont alors été conduites pour identifier et caractériser les usages de l'évaluation. La communauté de recherche en évaluation s'est accordée sur l'identification de trois principaux types d'usages de l'évaluation, et sur le fait que non seulement les rapports d'évaluation, mais aussi le processus de l'évaluation en tant que tel sont susceptibles d'être sources de changements. Les trois types d'usages sont :

- Le type instrumental : il correspond à la vision classique du rôle de l'évaluation, c'est-à-dire l'utilisation des résultats par les responsables politiques afin de modifier l'intervention publique;
- Le type conceptuel/apport d'un éclairage nouveau (*illumination*) : il est utilisé lorsque les résultats de l'évaluation contribuent à voir l'intervention publique sous un angle nouveau et à apporter des connaissances sur un phénomène potentiellement transposable à d'autres contextes;
- Le type persuasif : l'évaluation est utilisée pour légitimer une décision dans le cadre d'un argumentaire politique. Elle est parfois symbolique lorsque la conduite même de l'évaluation indépendamment de ses résultats suffit à appuyer une posture politique avant la conduite même de l'évaluation.

Le texte de Marvin Alkin et Sandy Taut (**texte 1**) synthétise ces travaux et présente ces différentes catégories d'usages de l'évaluation. Pour autant, si la variété des usages fait consensus auprès des évaluatrices et évaluateurs, des désaccords persistent quant à l'impact de l'évaluation sur de potentielles transformations. Ici deux positions s'opposent :

- Celle de Michael Patton, qui estime que l'évaluateur ou l'évaluatrice peut jouer un rôle central et déterminant dans l'utilisation des résultats de l'évaluation et en faire ainsi un outil de transformation majeure de l'action publique (**texte 2**). Ces éléments sont approfondis en détail dans la partie Utilité du présent livre.
- Celle de Carol H. Weiss, qui, s'appuyant sur sa connaissance du fonctionnement du champ politique, reconnaît que l'évaluation aussi rigoureuse soit elle n'est qu'une source d'information parmi d'autres, et que son impact est limité par le poids des autres facteurs d'influence sur la décision politique (**texte 3**).

Controverses : l'évaluation, influence parmi d'autres ou base de décision?

À la suite des premiers travaux sur l'identification des différents usages de l'évaluation, un champ de recherche spécifique a été dédié à l'approfondissement des connaissances à ce sujet. Une des premières démarches consiste à contester l'emploi du terme « utilisation/utilité » en tant qu'il ne permet pas d'appréhender les applications de l'évaluation qui ne sont pas basées sur l'utilisation des résultats, sur les effets non intentionnels, ni sur l'émergence graduelle d'impact dans le temps. Karen Kirkhart (**texte 4**) nous enjoint d'aborder ce questionnement en nous intéressant à « l'influence » de l'évaluation définie comme « la capacité ou pouvoir de choses ou de personnes à produire des effets sur d'autres par des moyens intangibles ou indirects » (Kirkhart, 2000 : 7).

Cette approche replace l'évaluation comme une contribution parmi d'autres dans le processus complexe de décision, et s'écarte d'une vision linéaire de son utilisation. Elle propose alors une catégorisation de l'influence de l'évaluation en trois dimensions :

- Celle de l'intention : l'usage est-il issu d'une démarche volontaire ou involontaire de l'évaluateur?
- Celle de la source : l'usage est-il issu du processus de l'évaluation ou de ses résultats?
- Celle de la temporalité : l'usage prend-il effet pendant l'évaluation, dès la fin de l'évaluation ou se produit-il sur le long terme?

Cette typologie présente l'avantage de proposer une conception intégrée qui permet d'englober l'ensemble des différents types d'influence de l'évaluation.

S'appuyant sur cette conception de l'influence de l'évaluation, Gary Henry et Melvin Mark (**texte 5**) se sont engagés dans l'investigation des mécanismes qui sous-tendent l'influence de l'évaluation. En d'autres

termes, les auteurs se demandent pourquoi, comment, dans quels cas, et auprès de qui l'évaluation peut exercer une influence. En utilisant les mêmes outils d'analyse que ceux déployés par un évaluateur ou une évaluatrice pour l'étude d'un dispositif, ils se sont efforcés d'identifier les leviers, et les étapes du processus expliquant les différents impacts des évaluations, ce qui les a conduits à formuler une « théorie de programme de l'évaluation ». Ainsi, Henry et Mark identifient-ils trois niveaux d'influence : individuel, interpersonnel et collectif, chacun opérant au moyen de mécanismes variés. Les auteurs poursuivent deux objectifs, d'une part contribuer à structurer les recherches sur l'influence de l'évaluation, et d'autre part, amener les évaluateurs et évaluatrices à formuler un plan de maximisation de la portée de leurs évaluations en activant les mécanismes qu'ils ont identifiés (Mark et Henry, 2004).

D'autres courants appellent à une institutionnalisation de l'évaluation comme source principale d'information de la décision publique (*evidence-based policy*) (voir le texte de H. White dans la partie Valeurs). D'un côté, ce mouvement enjoint les institutions politiques à s'appuyer de manière systématique sur des résultats probants issus de travaux de recherche ou d'évaluation; de l'autre, il pousse les évaluatrices et évaluateurs à avoir un rôle proactif de capitalisation des connaissances par l'identification et la formulation de bonnes pratiques (*best practices*) ou de leçons apprises (*lessons learned*) (Milton, 2010). Ce mouvement se retrouve notamment dans le développement des *What Works Centres* recensant par thématiques ou types de publics les leçons issues de la recherche ou de l'évaluation (Allard et Rickey, 2017).

Perspectives : repenser le rôle et les valeurs de l'évaluation pour la rendre plus utile?

Les réflexions actuelles autour de l'utilité et de l'influence de l'évaluation ont élargi la focale d'analyse pour s'intéresser, plus globalement, au rôle de l'évaluation dans l'état actuel du monde. En effet, s'intéresser à l'impact de l'évaluation dans le cadre d'un programme public, ou à l'échelle d'un projet, ne constitue qu'un effet intermédiaire de l'évaluation. L'objectif final est la conduite effective de meilleures politiques publiques, qui soient à même d'améliorer le bien-être des citoyens et des citoyennes (pour un développement de ce point, voir le texte de Sandra Mathison dans la partie Évaluatrice).

Dès lors, comment conduire une évaluation de sorte qu'elle embrasse le monde dans sa complexité, et qu'elle s'adapte aux différents contextes, enjeux et systèmes? Thomas Schwandt **(texte 6**) propose de penser une évaluation « post-normale » qui s'organiserait autour de principes-clés permettant d'élargir et d'approfondir le potentiel d'utilisation des évaluations. Il enjoint les équipes d'évaluation à penser les citoyen-ne-s dans une dynamique de coproduction avec les dispositifs publics, et engagé-e-s dans un questionnement éthique sur les limites et les présupposés de l'évaluation. Pour ce faire, il propose des approches posant la résilience comme but de l'action publique, favorisant la participation des citoyen-ne-s à la démocratie, en délaissant le raisonnement scientifique au profit d'un raisonnement pratique orienté vers l'action.

Pour finir, peut-on considérer que dans le contexte politique actuel de post-vérité (*post-truth*) l'évaluation a toujours un rôle à jouer? Quelle(s) réponse(s) est-elle en mesure d'apporter? Selon Robert Picciotto (**texte 7**) l'évaluateur ou l'évaluatrice doit impérativement s'engager dans une lutte contre les discours de post-vérité. Il lui est alors nécessaire d'initier une transformation de son métier. Cela requiert d'identifier les nouveaux buts des politiques publiques, leurs nouveaux objets, leurs nouveaux acteurs

(fondations, ONG), de s'internationaliser, de se diversifier, de se digitaliser etc. Cette transformation des techniques doit aller de pair avec une promotion explicite des valeurs démocratiques dans l'évaluation, et notamment du triptyque que propose E. House (House et Howe 1999) (voir ce texte dans la partie Évaluatrice) : l'inclusion (travailler avec des groupes sous-représentés), le dialogue (amener les parties prenantes à se comprendre), et la délibération (échanger au sujet des notions de valeur et de résultat).

Bibliographie

Alkin, Marvin C., et Jean A. King. 2017. « Definitions of Evaluation Use and Misuse, Evaluation Influence, and Factors Affecting Use ». *American Journal of Evaluation* 38(3) : 434-50. doi : 10.1177/1098214017717015.

Allard, Caroline et Rickey, Ben. 2017. *What Works Centres britanniques : quels enseignements pour des politiques fondées sur la preuve en France?* Paris : Agence nationale des solidarités actives.

Campbell, Donald T. 1969. « Reforms as Experiments ». *American Psychologist* 24(4) : 409-29.

Cousins, J. B. 2004. « Commentary : Minimizing Evaluation Misuse as Principled Practice ». *American Journal of Evaluation* 25(3) : 391-97.

Duflo, Esther. 2005. « Évaluer l'impact des programmes d'aide au développement : le rôle des évaluations par assignation aléatoire ». *Revue d'économie du développement* 19(2) : 185-226.

House, Ernest R. et Kenneth R. Howe. 1999. *Values in Evaluation and Social Research*. 1re éd. Thousand Oaks: Sage Publications.

Kirkhart, Karen E. 2000. « Reconceptualizing Evaluation Use: An Integrated Theory of Influence ». *New Directions for Evaluation* 2000(88) : 5-23. doi : 10.1002/ev.1188.

Mark, Melvin M., et Gary T. Henry. 2004. « The Mechanisms and Outcomes of Evaluation Influence ». *Evaluation* 10(1) : 35-57. doi : 10.1177/1356389004042326.

Milton, N. J. 2010. *The lessons learned handbook: practical approaches to learning from experience*. Oxford, UK: Chandos Publishing.

Patton, Michael Quinn. 1988. « Reports on Topic Areas: The Evaluator's Responsibility for Utilization ». *American Journal of Evaluation* 9(2).

Rajkotia, Yogesh. 2018. « Beware of the Success Cartel: A Plea for Rational Progress in Global Health ». *BMJ Global Health* 3(6). doi : 10.1136/bmjgh-2018-001197.

Spenlehauer, Vincent. 2016. « La (f)utilité gouvernementale de l'évaluation des politiques publiques, quelques leçons américaines et françaises ». *LIEPP Working Paper* 49.

Viveret, Patrick. 1989. *L'évaluation des politiques et des actions publiques. Propositions en vue de l'évaluation du Revenu minimum d'insertion.* Paris : Rapport au Premier Ministre. Commissariat au Plan.

Weiss, Carol. H. 1988. « If Program Decisions Hinged Only on Information: A Response to Patton ». *Evaluation Practice/American Journal of Evaluation* 9(3) : 15-28. doi : 10.1177/109821408800900302.

Weiss, Carol H. 1972. « Utilization of evaluation: Toward comparative study ». in *Evaluating action programs: Readings in social action and education.* Boston: Allyn & Bacon.

1. Démêler les usages de l'évaluation

MARVIN C. ALKIN ET SANDY M. TAUT

[Traduit de : Alkin, Marvin C., Sandy M. Taut. 2003. « Unbundling evaluation use ». *Studies in Educational Evaluation*, 29 : 1-12 (Extraits). Traduction de Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation d'Elsevier.]

[…] De toute évidence, les connaissances issues d'une évaluation sont le produit de la manière dont cette évaluation a été conduite. Celles-ci peuvent être générées non seulement à la fin de l'évaluation, mais aussi à de nombreux moments au fil de son déroulement. Par exemple, l'évaluation d'un programme de prévention du VIH/SIDA chez les jeunes peut produire des connaissances non seulement en diffusant un rapport à la fin de la première année de mise en œuvre du programme, mais aussi en amenant les opérateurs et opératrices du programme à réfléchir à leurs pratiques par exemple en sollicitant leurs avis dans l'élaboration d'un questionnaire d'enquête, ou lors de conversations informelles sur la théorie sous-jacente du programme.

En outre, les connaissances issues de l'évaluation sont produites en relation avec un programme particulier. Ainsi, bien que l'évaluation en tant que telle détermine la nature des connaissances produites en matière d'évaluation, d'autres facteurs entrent également en jeu – et tous ces facteurs interagissent les uns avec les autres. Les caractéristiques des utilisatrices et utilisateurs (les personnes impliquées) et les facteurs contextuels (la situation ou le contexte environnant le programme évalué) ont également un impact sur les connaissances produites par l'évaluation. Des recherches antérieures ont montré que ces mêmes caractéristiques influencent l'utilisation des résultats de l'évaluation. Dans un examen systématique des facteurs affectant l'utilisation de l'évaluation, Alkin (1985) établit une distinction entre les « facteurs d'évaluation », les « facteurs humains » et les « facteurs contextuels ».

Voyons d'abord comment les caractéristiques de l'évaluation en tant que telles (facteurs d'évaluation) façonnent la production de connaissances issues d'une évaluation. Tout d'abord, il faut s'intéresser à la démarche d'évaluation. Quelle approche générale a été mobilisée? Y a-t-il eu une approche comparative? Quels types d'outils ont été utilisés pour la collecte des données? Quelles logiques d'analyse ou d'agrégation ont été utilisées? Deuxièmement, il faut s'intéresser à la transmission d'informations, pendant la conduite de l'évaluation et une fois qu'elle est finalisée. Troisièmement, il existe plusieurs niveaux d'intensité d'échanges d'informations qui peuvent avoir lieu au cours de l'évaluation entre l'évaluateur et les utilisateurs potentiels (parties prenantes). Quatrièmement, la manière dont les connaissances issues de l'évaluation sont présentées détermine également la nature des connaissances produites. Enfin, les caractéristiques personnelles et les actions de l'évaluateur ou évaluatrice sont également à prendre en compte. Notamment son engagement pour favoriser l'utilisation, sa volonté d'impliquer les potentiels utilisateurs, sa connaissance du champ politique, sa crédibilité, sa relation avec les potentiel-le-s utilisateurs et utilisatrices, etc.

Le type de connaissances qu'une évaluation produit et la façon dont ces connaissances sont utilisées dépendent également du contexte particulier du programme (facteurs contextuels). Les facteurs contextuels regroupent des caractéristiques relatives aux contraintes de l'évaluation, qu'elles soient contractuelles ou budgétaires, des caractéristiques inter et intra-organisationnelles, ainsi que les facteurs relatifs à la communauté environnante du programme. Les caractéristiques du projet ou du programme, par exemple sa maturité ou son ancienneté, relèvent aussi des facteurs contextuels.

Les facteurs humains sont caractérisés par Alkin (1985) comme relevant à la fois des caractéristiques de l'évaluateur/-trice et de l'utilisateur/-trice. Nous avons choisi ici de considérer les caractéristiques de l'équipe d'évaluation comme faisant partie des facteurs d'évaluation et nous nous concentrerons sur les caractéristiques de l'utilisateur ou de l'utilisatrice

comme troisième élément majeur. Les caractéristiques de l'utilisateur/-trice recouvrent des éléments tels que son identité, sa responsabilité organisationnelle et diverses caractéristiques personnelles et professionnelles. La caractéristique la plus importante restant l'intérêt des utilisateurs et utilisatrices potentiel-le-s pour l'évaluation en général (et pour cette évaluation en particulier), ainsi que leur engagement à en utiliser les enseignements.

Des synthèses de recherches sur les facteurs associés à l'utilisation de l'évaluation ont été réalisées par Cousins et Leithwood (1986), Shulha et Cousins (1997) ainsi qu'Hofstetter et Alkin (2003). Toutefois, ce vaste corpus de recherches s'est concentré sur l'utilisation des résultats de l'évaluation et n'a pas examiné de manière approfondie la façon dont le processus d'évaluation peut aussi être utilisé. Ainsi, les principales recherches sur l'utilisation de l'évaluation ont essentiellement porté sur la question suivante : comment la connaissance produite par l'évaluation (ses conclusions) a-t-elle été convertie en ce que l'on appelle une utilisation instrumentale ou conceptuelle?

Cette distinction entre l'utilisation instrumentale ou conceptuelle des résultats a été suggérée pour la première fois par Rich (1977). L'utilisation instrumentale désigne les cas où les connaissances ont été utilisées pour influer de manière directe sur une action, comme la prise de décisions particulières au sujet d'un programme. L'utilisation conceptuelle (parfois appelée « utilisation éclairée » (*enlighted*)) décrit des cas où les résultats de l'évaluation n'ont pas directement été utilisés pour une prise de décision, mais ont contribué à modifier la manière dont les utilisatrices et utilisateurs appréhendent ou conceptualisent certains aspects du programme évalué. Le concept d'utilisation symbolique, également répandu dans la littérature sur l'évaluation, décrit des situations où l'évaluation est utilisée pour justifier une décision préalable ou pour démontrer qu'un programme est prêt à être évalué afin d'améliorer la réputation d'un gestionnaire de programme ou d'un décideur. Owen (1999) distingue ces deux types d'utilisation d'évaluation, celle de la justification d'une décision préalable d'une part et celle de l'amélioration

de la réputation d'autre part, en qualifiant la première de « légitimatrice ». Le terme « d'évaluation symbolique » est réservé aux cas où l'évaluation est conduite uniquement au service d'un renforcement de statut d'une personne ou comme acte symbolique.

Plus récemment, le concept d'utilisation du processus d'évaluation a pris de l'importance, déplaçant la focale de l'utilisation des résultats de l'évaluation à la façon dont la conduite de l'évaluation (le processus de l'évaluation) a des répercussions sur les personnes ou les organisations. Patton (1997 : 90) définit l'utilisation du processus d'évaluation comme « des changements individuels dans la pensée et le comportement, ainsi que des changements de programme ou d'organisation dans les procédures et la culture, qui surviennent chez les personnes qui participent à l'évaluation à la suite de l'apprentissage qui se produit au cours du processus d'évaluation » (Patton, 1997 : 90).

Par exemple, on est en présence d'utilisation du processus d'évaluation si le programme est modifié en raison du processus de réflexion que l'évaluation organise, plutôt que comme une conséquence des conclusions qu'elle a produite. D'autres, comme Preskill et Torres (2000), se concentrent plus particulièrement sur l'étude des effets d'apprentissage que le processus d'évaluation est susceptible de générer au niveau tant individuel qu'organisationnel.

Le processus d'évaluation produit-il des connaissances? Le cas échéant, en quoi l'utilisation de connaissances produites par le processus d'évaluation se distingue-t-elle de l'utilisation de connaissances issues des résultats de l'évaluation? L'utilisation du processus de l'évaluation, contrairement à l'utilisation des résultats de l'évaluation, ne se limite pas à un apprentissage basé sur l'utilisation (ou non) des connaissances produites par l'évaluation. En consultant la littérature relative à la psychologie sociale (Fischer et Wiswede 1997; Aronson, Wilson et Akert 1999), nous avons constaté que l'apprentissage comporte deux composantes qui interagissent entre elles : l'acquisition et l'accumulation de connaissances, d'une part, l'acquisition et la modification du

comportement, d'autre part. Dans le cas de l'utilisation des résultats, le premier type d'apprentissage domine. Cependant, dans le cas de l'utilisation du processus, la conduite de l'évaluation en tant que telle permet également aux utilisateurs potentiels d'acquérir de nouvelles compétences et de modifier leur comportement. Par exemple, le personnel d'un programme de lutte contre la toxicomanie pourrait, au cours de l'évaluation (en raison des questions posées par l'évaluateur ou l'évaluatrice et de la prise de recul introduite par l'évaluation), avoir l'occasion de réfléchir à ses interactions avec ses publics. Ils et elles pourraient détecter certains modèles de comportement qui s'avèrent inefficaces, ce qui entraînerait un remodelage de ces schémas d'interaction au cours du déploiement de l'évaluation. Nous supposons que la manière dont l'évaluation est menée, le contexte et les utilisateurs/-trices (tel-le-s que décrit-e-s ci-dessus) ont un impact sur la possibilité et la manière dont l'utilisation du processus d'évaluation peut advenir.

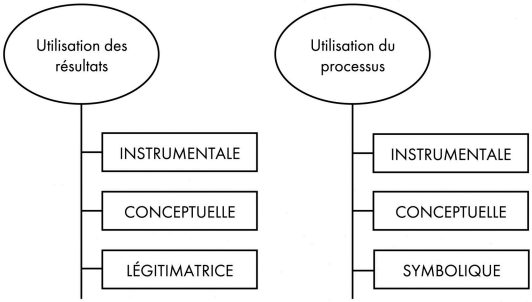
Clarifier les types d'utilisations d'évaluation

Nous avons déjà évoqué les distinctions entre l'utilisation des résultats et l'utilisation du processus d'évaluation. En général, l'utilisation des résultats a été examinée en fonction des catégories instrumentale, conceptuelle et symbolique (ou légitimatrice). Ces concepts sont assez bien compris lorsque l'on considère l'utilisation des résultats (voir ci-dessus). Nous maintenons que l'utilisation du processus d'évaluation n'est pas une autre catégorie au même titre que ces types d'utilisation, mais plutôt un autre domaine d'utilisation. Autrement dit, les résultats peuvent être utilisés, ou le processus peut être utilisé (voir la figure 3). En outre, l'utilisation du processus peut se faire de façon instrumentale ou conceptuelle. Ainsi, un changement instrumental peut être effectué, ou une décision peut être prise en s'appuyant sur un processus d'évaluation. Par exemple, les questions posées par l'évaluateur ou l'évaluatrice sur la nature du programme peuvent conduire à des reconsidérations, des

réexamens et des modifications du programme. Nous pensons qu'il s'agit là d'une utilisation instrumentale du processus d'évaluation. L'utilisation du processus d'évaluation peut également se faire sur le plan conceptuel. Le processus d'évaluation peut contribuer à modifier l'attitude des utilisateurs et utilisatrices à l'égard de l'importance de l'évaluation ou du rôle potentiel de différentes parties prenantes dans la prise de décision. Greene (1988 : 111) décrit ces « utilisations conceptuelles du processus d'évaluation ». Elle estime qu'un processus participatif d'évaluation accroît la probabilité d'utilisation des résultats de l'évaluation via une utilisation conceptuelle du processus d'évaluation qui contribue en tant que telle à favoriser « les dispositions pour l'utilisation » des parties prenantes.

L'utilisation symbolique et l'utilisation légitimatrice suggérée par Owen (1999), constituent une autre distinction intéressante. L'utilisation légitimatrice, comme suggérée, fait référence à la légitimation de décisions préalables. On peut supposer qu'il s'agit de résultats d'évaluation démontrant qu'une décision prise antérieurement était bien fondée. Par conséquent, l'utilisation légitimatrice fait clairement partie de la catégorie d'utilisation des résultats. D'autre part, l'utilisation symbolique fait référence aux processus d'évaluation en tant qu'actes symboliques qui pourraient renforcer la position du décideur, du programme ou autre entité. L'utilisation symbolique fait donc spécifiquement référence à un type de légitimation par l'engagement dans un processus sans particulièrement tenir compte des résultats. Il nous semble donc raisonnable de différencier l'utilisation légitimatrice ayant trait à l'utilisation des résultats et l'utilisation symbolique ayant trait à l'utilisation des processus.

La figure suivante résume ces distinctions :



[…]

Bibliographie

Alkin, Marvin C. 1985. A *guide for evaluation decision makers*. Beverly Hills: Sage Publications.

Aronson, Elliott, Timothy D. Wilson et Robin M. Akert. 1999. *Social psychology*. 3e éd. Englewood Cliffs: Prentice-Hall.

Cousins, J. Bradley, et Kenneth A. Leithwood. 1986. « Current Empirical Research on Evaluation Utilization ». *Review of Educational Research* 56(3) : 331-64.

Fischer, Lorenz, et Günter Wiswede. 1997. *Grundlagen der Sozialpsychologie [Foundations of Social Psychology]*. Munich : Oldenbourg.

Greene, Jennifer G. 1988. « Stakeholder participation and utilization in program evaluation ». *Evaluation Review* 12 (2) : 91-116. doi : https://doi.org/10.1177/0193841X8801200201

Hofstetter, Carolyn H. et Marvin C. Alkin. 2003. « Evaluation use revisited ». in *International handbook of educational evaluation*, édité par D. L. Stufflebeam, T. Kellaghan et L. Wingate. Boston: Kluwer International Handbooks of Education, p. 197-222.

Owen, John M. 1999. *Program evaluation: Forms and approaches*. 2e éd. Londres : Sage Publications.

Patton, Michael Q. 1997. *Utilization-focused evaluation: The new century text*. 3e éd. Thousand Oaks: Sage Publications.

Preskill, Hallie, et Rosalie T. Torres. 2000. « The learning dimension of evaluation use ». *New directions for evaluation* 88 : 25-37.

Rich, Robert F. 1977. « Use of Social Sciences Information by Federal bureaucrats: Knowledge for Action Versus Knowledge for Understanding ». in *Using Social Research in Public Policy Making*, édité par C. H. Weiss. Lexington, Mass: Heath.

Shulha, Lyn M., et J. Bradley Cousins. 1997. « Evaluation use: Theory, research, and practice since 1986 ». *Evaluation Practice* 18(3) : 195-208. doi : https ://doi.org/10.1177%2F109821409701800302

2. La responsabilité de l'évaluateur quant à l'utilisation de l'évaluation

MICHAEL Q. PATTON

[Traduit de : Patton, Michael Q. 1988. « Reports on Topic Areas: The Evaluator's Responsibility for Utilization ». *American Journal of Evaluation*, 9(2) : 5-24 (Extraits). Traduction par Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

[...] L'évaluation a vocation à améliorer les programmes, à accroître leur efficacité et à fournir des informations utiles à la prise de décision. Telle est l'utilisation de l'évaluation. Telle est notre responsabilité.

L'utilisation escomptée par les utilisateurs et utilisatrices ciblé-e-s

Permettez-moi de définir ce que j'entends par utilisation du point de vue de la responsabilité. Ceux et celles d'entre vous qui connaissent la première édition d'*Utilisation-Focused Evaluation* (Patton, 1978) savent qu'une des faiblesses de ce livre a été mon échec à définir véritablement ce qu'est l'utilisation. La deuxième édition (Patton, 1986) corrige cette lacune. Par utilisation, j'entends l'utilisation escomptée de la part des utilisateurs et utilisatrices ciblé-e-s. Si vous y réfléchissez, cela ressemble fortement à la manière de formuler un objectif de programme, mais ici pour une évaluation. L'utilisation ainsi escomptée s'inscrit dans une approche par les résultats. Cela signifie que nous négocions en amont avec les utilisatrices et utilisateurs ciblé-e-s ce qu'une évaluation doit

réaliser pour qu'elle apporte plus qu'elle ne coûte. Ensuite, on attend de nous que nous soyons à la hauteur, ou que nous soyons jugé-e-s comme n'ayant pas atteint notre objectif d'utilisation.

L'approche par les résultats est à la mode aujourd'hui dans le domaine du management, de l'élaboration de programmes et nous devrions l'appliquer – et je crois que nous sommes en train de le faire – à l'évaluation.

[…] Il ne suffit pas d'avoir les méthodes d'évaluation, les attributs d'un-e évaluateur/-trice, d'avoir des institutions dédiées, d'écrire et de publier à ce sujet. J'estime que l'on n'est pas évaluateur/-trice tant que l'on n'a pas obtenu que ses résultats soient utilisés, [au sens de] l'utilisation escomptée par les utilisateurs et utilisatrices ciblé-e-s.

Cela peut sembler effrayant pour celles et ceux d'entre vous qui n'ont pas eu affaire sérieusement à la question de la responsabilité en matière d'utilisation. Mais la bonne nouvelle est que nous savons comment atteindre des niveaux élevés d'utilisation et […] beaucoup y parviennent.

L'atteinte d'un niveau d'utilisation satisfaisant découle de notre connaissance de la manière dont fonctionnent les politiques publiques. Cette connaissance comprend la capacité à proposer une approche adaptée pour interagir avec les utilisateurs et utilisatrices ciblé-e-s et les principales parties prenantes afin de les former en tant qu'utilisateurs et utilisatrices de l'information, et de travailler avec elles et eux sur un engagement mutuel pour l'utilisation du processus d'évaluation et de ses résultats. […]. Je voudrais discuter [ci-dessous] de certaines stratégies que j'ai trouvées utiles pour assumer notre responsabilité en veillant à ce que les utilisateurs et utilisatrices ciblé-e-s utilisent l'évaluation de la façon escomptée.

Lever les appréhensions

Parfois, accroître l'utilisation passe d'abord par lever les appréhensions qui l'entourent. Je fais ici allusion à une situation, que je rencontre souvent, où les équipes sont hostiles à l'évaluation et/ou en ont peur. La peur est très présente dans notre travail. (Vous pouvez ressentir une partie de cette peur vous-même lorsque vous pensez à ce que cela signifie d'être évalué-e sur la base de ce que vous faites en tant qu'évaluateur/-trice, d'être tenu-e responsable de l'utilisation selon les lignes décrites ci-dessus.)

Je m'efforce, pour ma part, de faire face à cette peur de l'évaluation dès le départ. Je réunis les principaux utilisateurs et principales utilisatrices, et des représentant-e-s de l'administration, des financeurs/-ceuses, de l'équipe et des client-e-s pour une première réunion au cours de laquelle nous présentons la forme que prendra le processus d'évaluation, et où nous commençons à susciter un engagement à l'utilisation. Cela me donne l'occasion d'évoquer leur image de l'évaluation et des évaluateurs/-trices. Je trouve que c'est une discussion essentielle pour comprendre leurs points de vue.

Au cours d'une telle discussion, nous avons l'occasion et la responsabilité de contrôler l'image que nous projetons de l'évaluation. Chaque fois que nous faisons une évaluation, nous projetons une image de la profession. Chacun-e d'entre vous[1] représente la profession et projette une image de la profession et de ce que nous avons à offrir. Nous établissons qui nous sommes par ce que nous faisons et par les résultats que nous obtenons. Ainsi, une approche en vue d'accroître l'utilisation consiste à créer dès le début d'un processus d'évaluation une attente positive quant au fait que cette évaluation sera utile. Une telle perspective positive quant à l'utilité peut nécessiter une rupture avec les expériences passées des membres

1. Membres de la Société Américaine d'Évaluation.

de l'équipe en matière d'évaluation, de sorte qu'ils et elles peuvent être sceptiques, et ont le droit de l'être jusqu'à ce que nous leur prouvions le contraire. Mais il est important que nous placions en nous-mêmes et ces personnes une attente positive et responsable quant au fait que l'évaluation devrait être et peut effectivement être utile.

Poser les bonnes questions

Un deuxième élément à prendre en considération au moment où nous entamons ce dialogue avec les utilisatrices et utilisateurs ciblé-e-s est de poser les bonnes questions. Cela demande beaucoup d'habileté. Qu'il s'agisse de procéder à une étude d'évaluabilité – ce qui est un moyen de comprendre ce qu'il est possible de faire dans le cadre de l'évaluation – ou d'interagir avec une ou plusieurs parties prenantes, je pense que poser les bonnes questions c'est notamment s'interroger sur l'utilisation escomptée par les utilisateurs et utilisatrices ciblé-e-s et les éléments qui pourraient en témoigner. Comment pourrions-nous observer l'utilité d'une évaluation pour améliorer un programme et la prise de décision dans un contexte en particulier?

Apprendre à poser des questions et à réellement écouter ce que disent les commanditaires est une compétence essentielle. Jeri Nowakowsi me parlait de son article dans le prochain volume de *New Directions* consacré aux perceptions des commanditaires. Elle a dit que si elle devait le réduire à un seul argument ultime, ce serait que les relations les plus efficaces entre évaluateurs/-trices et commanditaires ont été celles où les évaluateurs/-trices ont vraiment écouté leurs commanditaires et travaillé avec elles et eux pour faciliter l'utilisation.

Dans les formations à la conduite d'entretien que je réalise, poser les bonnes questions est un élément majeur pour les activités de conseil ou de collecte de données. Poser de bonnes questions n'est pas toujours facile. L'une des façons dont j'aime former les enquêteurs et enquêtrices à

poser de bonnes questions est de leur faire mener des entretiens avec des enfants. Ce qui est merveilleux avec les enfants, c'est qu'ils et elles n'ont pas appris toutes les subtilités pour prétendre que votre question n'était pas idiote alors qu'elle l'est. Les enfants ont tendance à répondre à tout ce qui leur est demandé, ce qui peut être assez embarrassant.

L'une des premières évaluations que j'ai faites concernait des programmes d'éducation alternative (*open education*) dans le Dakota du Nord. Les défenseurs et défenseuses de l'éducation alternative affirmaient que l'apprentissage était très amusant pour les élèves parce qu'il les plaçait en son centre. Ils ont même laissé entendre que l'apprentissage était tellement amusant que les enfants ne s'occupaient pas de choses comme la récréation, car aller jouer à l'extérieur n'était pas mieux que ce qui se passait en classe. Les enfants continuaient donc à faire ce qu'ils et elles faisaient en classe. Par conséquent, nous avons inclus dans nos entretiens avec les enfants une question sur ce qu'ils et elles faisaient pendant la récréation.

Lors de mon troisième ou quatrième entretien, j'interrogeais une petite fille de première année, et j'en suis arrivé à cette question cinq à dix minutes après le début de l'entretien.

Je lui ai demandé : « Que fais-tu pendant la récréation? » Elle m'a répondu : « Je vais dehors et je joue sur les balançoires de la cour de récréation ». Je lui ai dit : « Pourquoi vas-tu dehors? » Elle m'a regardé d'un air perplexe et m'a répondu : « Parce que c'est là que sont les balançoires ». Et je pouvais voir qu'elle savait qu'elle avait sur les bras un de ces adultes qui ne comprenait pas vraiment comment les choses fonctionnent, que si vous voulez vous balancer sur les balançoires, vous devez aller dehors, là où sont les balançoires. Elle m'a donc très patiemment expliqué la nature du monde, puis nous avons pu passer à autre chose.

Il y a des moments où nous devons poser des questions qui semblent stupides. Peter Drucker, le conseiller en gestion, dit que c'est ce qu'il fait. Il rencontre des conseils d'administration d'entreprises et commence par leur demander : « Quel est votre secteur d'activité? ». Il rapporte que ses clients et clientes sont souvent agacé-e-s, si elles et ils n'ont pas été prévenu-e-s, parce qu'ils et elles pensent qu'il n'a pas fait ses devoirs. « Comment ce consultant onéreux peut-il se présenter et nous demander quel est notre secteur d'activité? ». Drucker travaille ensuite avec eux pendant plusieurs jours, les aidant à comprendre ce qu'est réellement leur entreprise, car ils et elles ne le savent généralement pas. Et le fait de le découvrir est source de changement pour cette entreprise.

C'est en grande partie ce que nous faisons : poser des questions pour aider l'action publique à comprendre ce qu'elle fait réellement, vérifier si elle l'a bien fait afin de faire mieux. Dans ce processus, ce sont à la fois les acteurs et actrices de l'action publique et nous-mêmes qui sommes responsables.

S'adapter à la situation en tant qu'évaluateur expert

En plus de poser les bonnes questions, il faut apprendre à analyser à la situation et à reconnaître les variations dans les situations et les personnes afin d'adapter son processus à celles-ci. Les évaluateurs et évaluatrices expérimenté-e-s se sont perfectionné-e-s dans la reconnaissance des situations. Le partage d'expertise est l'un des objectifs de la présente réunion de l'AEA (American Evaluation Association). Je voudrais consacrer quelques minutes à présenter ma vision de ce que cela signifie d'être un évaluateur ou une évaluatrice expert-e, pour réfléchir à ce que vous pouvez accomplir grâce à votre expertise.

Il faut d'abord reconnaître que l'expertise ne vient pas du jour au lendemain et qu'elle ne résulte pas de la seule formation. L'expertise est le fruit d'heures de travail d'évaluation, mais cela vaut la peine de savoir

ce que ces heures peuvent donner si vous êtes prêt-e à travailler pour devenir un évaluateur ou une évaluatrice expert-e. Pour développer cette vision de l'expertise, je m'appuierai sur d'autres domaines qui se sont penchés plus en détail sur ce que signifie l'expertise, pour réfléchir aux implications pour l'évaluation.

J'ai trouvé un article intéressant qui fait état d'études sur l'expertise dans différents jeux – les jeux d'échecs de classe mondiale, le sommet du domaine. Permettez-moi de lire un extrait du travail d'Etheredge (1979) sur l'expertise. Il dit :

> Il faut au moins 15 ans de dur labeur pour que même les individus les plus talentueux et talentueuses deviennent des grands maîtres d'échecs internationaux : ce qu'ils et elles semblent apprendre est un répertoire pour reconnaître les types de situations ou de scénarios ou les sensibilités intuitives et comprendre comment ces situations vont se dérouler. Simon estime un répertoire différentiel de 50 000 reconnaissances de situations pour des échecs de classe mondiale. Il y a également une certaine augmentation de la capacité globale de planification stratégique à long terme – les débutantes et débutants ont généralement du mal à aller au-delà d'un coup d'avance; les grands maîtres d'échecs internationaux anticipent souvent 3 ou parfois 5 coups futurs en fonction des différentes réactions possibles à leurs coups (p.40).

Je suggère qu'il y ait un parallèle ici avec le fait d'anticiper l'utilisation et de savoir comment la faire advenir.

> Des données provenant de joueurs et joueuses d'échecs, de poker, de tennis et d'autres professionnel-le-s expérimenté-e-s et couronné-e-s de succès, suggèrent la théorie générale selon laquelle un autre enseignement de l'expérience est la capacité de diagnostiquer non seulement des situations de jeu spécifiques, mais aussi de modéliser différents adversaires (Etheredge, 1979 : 40).

Etheredge rapporte également qu'il est probable que les joueurs et joueuses expérimenté-e-s aient développé une analyse plus efficace et la capacité d'écarter les informations inutiles, et qu'ils et elles aient un sens approximatif, qui semble intuitif, mais qui est en fait un sens pratique, de ce sur quoi ils et elles doivent porter leur attention. À mon avis, vous aurez du mal à trouver une meilleure définition de ce qu'implique l'expertise en matière d'évaluation : c'est-à-dire l'identification d'une situation, de la réactivité, de l'anticipation et la capacité d'analyser les gens – en sachant où, quand et comment centrer son attention.

[…]

Réflexivité et évaluation

L'engagement à atteindre un niveau d'expert implique de réfléchir à notre propre pratique, d'appliquer nos compétences en matière d'évaluation à ce que nous faisons, comme le font le GAO (*General Accounting Office*), l'unité d'évaluation du FBI, ou encore de nombreux évaluateurs internes. Cela implique de prendre le temps d'étudier notre propre travail pour découvrir ce qui a fonctionné et ce qui n'a pas fonctionné, et de procéder à une évaluation formative de nos propres processus pour tirer des enseignements de ce que nous faisons.

Il est triste de constater que nous appliquons si peu nos propres compétences en matière d'évaluation à notre propre pratique. Fritz Steele, qui est consultant auprès des consultant-e-s, appelle cela « le dilemme action-réflexion ». Il dit que l'action pour les consultant-e-s consiste habituellement à aider les commanditaires à prendre le temps de réfléchir. Mais quand le ou la consultant-e prend-il ou elle le temps de conduire un travail réflexif sur ses expériences personnelles? Ce dilemme action-réflexion résulte, selon Steele (1975) de

> l'attractivité, sur le plan émotionnel, du volet 'action'. Il y a toujours la tentation de passer à une nouvelle action sans avoir complètement analysé ce qui s'est passé dans le cadre d'un projet terminé, sans avoir recueilli les commentaires des clients, examiné les notes et discuté de façon analytique de l'expérience avec les collègues. Prendre le temps d'évaluer un processus d'évaluation, d'examiner une expérience de conseil ou d'assurer le suivi de l'utilisation des résultats d'une étude sont des moyens de poursuivre l'apprentissage (p.19).

Tirer des leçons de votre apprentissage fait partie de l'engagement à devenir un professionnel ou une professionnelle compétent-e, et cela fait partie de ce que ce forum, et les réunions de l'Association Américaine d'Évaluation, sont censés faciliter. C'est une partie importante de la raison pour laquelle nous sommes ici : apprendre les uns des autres et apprendre à partir de cette réflexivité sur nos propres processus.

Défendre l'évaluation

Un dernier espoir que je voudrais vous laisser dans le cadre de cette discussion sur notre ambition et notre responsabilité est que nous devons être des défenseurs et défenseuses de l'évaluation. Nous pouvons et devons défendre l'évaluation, à partir du constat qu'il s'agit d'un produit et d'un processus de qualité qui peuvent améliorer les programmes.

Je voudrais utiliser ici une métaphore qui ne plaira peut-être pas à tous et toutes, compte tenu notre appétence pour la recherche et de certaines façons dont nous percevons notre rôle.

Lorsque je cherche à accroître mon expertise, je fais feu de tout bois. J'essaie de m'inspirer d'un certain nombre de domaines différents et des compréhensions que les gens ont développées dans des domaines plus anciens que le nôtre. La vente est l'un des domaines auxquels je me suis

intéressé. De très nombreuses recherches ont été menées sur ce qui fait d'une personne un vendeur ou une vendeuse efficace. Une grande partie de la formation à la vente n'est que du battage publicitaire, mais j'ai écouté des cassettes et lu ce que les meilleur-e-s vendeurs et vendeuses ont à dire sur l'efficacité. Je pense que nous pouvons en tirer des leçons. Je n'ai le temps d'en mentionner que deux d'entre elles.

L'une des leçons est tirée d'un consensus apparent dans la littérature sur la vente, à savoir qu'une condition préalable pour être un vendeur ou une vendeuse efficace est d'avoir un produit de qualité qui inspire la confiance et l'engagement. J'espère que mes remarques d'aujourd'hui vous auront fait comprendre que je crois fermement à l'évaluation en tant que processus et produit utile pour l'amélioration des programmes. Lorsque je parle aux responsables de programmes, je peux leur dire avec conviction que l'évaluation peut être une source de changement dans leur programme. Je peux leur assurer qu'un processus d'évaluation axé sur l'utilisation, entrepris avec délibération et sérieux, peut faire la différence dans ce qu'ils font. C'est ce que je crois. Je crois qu'il existe des preuves de cette affirmation, y compris des preuves provenant de nos réunions, de sorte que je puisse ainsi « vendre » l'évaluation et son utilité.

Toutefois, un produit de qualité ne semble pas suffire. L'habileté à vendre est nécessaire pour communiquer les vertus du produit. Cette compétence commerciale que je trouve particulièrement intrigante est apparemment un vrai problème pour les vendeurs et vendeuses en ce sens qu'elle est l'une des choses qu'ils et elles doivent travailler encore et encore pour la maîtriser. Le défi que cette compétence représente et constitue, je pense, une analogie intéressante avec nos propres difficultés à vendre l'évaluation. En ce qui concerne cette compétence, je voudrais citer l'un des meilleurs vendeurs, l'un des plus grands prestidigitateurs dans ce domaine, mais aussi un homme qui a une réelle connaissance de la vente et qui a les chiffres pour soutenir ses affirmations – Zig Ziglar. Laissez-moi vous dire ce que Zig Ziglar a dit dans un récent bulletin national destiné à sa profession. On lui a demandé des conseils pour augmenter les ventes et il a répondu que la plus grande faiblesse de la

plupart des vendeurs et vendeuses se situe au moment de conclure une vente. Il cite des études montrant que 63 % des entretiens de ventes se terminent sans que le vendeur demande explicitement au client ou à la cliente s'il ou elle souhaite passer commande. Il poursuit en notant que Herb True, de l'Université Notre Dame, a constaté que 46 % des vendeurs et vendeuses qu'il a interrogé-e-s ne posent la question qu'une seule fois, puis abandonnent; 24 % la posent deux fois; 14 % trois fois; et 12 % abandonnent après la quatrième tentative. Pourtant, ses recherches sur les ventes efficaces montrent que 60 % de toutes les ventes sont réalisées après la cinquième tentative de clôture (Ziglar, 1987).

Dès lors, celles et ceux d'entre vous qui, avec douceur, suggèrent une fois à un utilisateur ou une utilisatrice potentiel-le qu'il y a peut-être quelque chose d'important à faire avec l'évaluation, voient les implications de ma comparaison avec la vente. « Conclure » l'utilisation de l'évaluation, comme conclure une vente, est une chose à laquelle on travaille. Avoir la conclusion en tête, c'est avoir une vision claire de ce que signifie l'utilisation d'une évaluation, et ensuite rechercher cette utilisation, pas une fois, pas deux fois, pas trois fois, pas quatre fois, mais proposer cette utilisation aussi longtemps qu'il le faut pour vous acquitter de votre responsabilité et atteindre l'utilisation escomptée par les utilisateurs et utilisatrices ciblé-e-s.

L'une des façons d'y parvenir est de visualiser la « clôture » de l'évaluation dès le début, de négocier une compréhension commune de ce que cela signifie de clore l'évaluation, c'est-à-dire d'atteindre une utilisation effective. Vous devez communiquer sur votre engagement envers l'utilité à chaque étape de l'évaluation.

L'une des façons dont je porte cet engagement envers l'utilisation au cours des premières négociations est de demander s'ils s'attendent à un rapport final. Cette question semble attirer l'attention des gens. « Voulez-vous un rapport final? ». Ils me regardent et disent : « Pardon? ». Et je dis : « Voulez-vous un rapport final? ». Ils disent : « Bien sûr, c'est pour cela que nous faisons ça, pour obtenir un rapport. ».

Et je réponds : « Non, nous faisons cela pour obtenir des informations qui vous permettront d'améliorer la conception de votre programme et votre prise de décision. Un rapport écrit final est un moyen de communiquer les informations que nous obtenons. Mais il y a des preuves substantielles montrant que ce n'est pas toujours le moyen le plus efficace. Il est très coûteux de rédiger des rapports finaux. Il existe d'autres façons de traiter ces informations. Parlons donc de ce que vous souhaitez obtenir pour votre argent, en termes d'utilisation. C'est ce que nous essayons vraiment d'obtenir ici. Pas un rapport, mais l'utilisation. Parlons des moyens les plus efficaces d'obtenir une utilisation et voyons si un rapport final, écrit et coûteux est le moyen d'y parvenir ».

Je constate souvent que c'est à ce moment-là qu'ils commencent enfin à comprendre que ce dont je parle diffère de la simple production d'un épais rapport d'évaluation pour pouvoir classer le programme sous la rubrique « a été évalué ».

Une mission et une vision

Je pense donc que nous avons devant nous, dans les standards de l'évaluation, dans le travail de cette réunion, et dans les exemples que j'ai cités, la vision d'une approche positive et proactive de l'évaluation axée sur l'utilisation qui soit source de changement et qui soit responsable. Le processus d'évaluation est source de changement au fur et à mesure de son déroulement. Les résultats de l'évaluation sont source de changement lorsqu'ils sont utilisés de la manière escomptée par les utilisateurs et les utilisatrices ciblé-e-s.

Les nouveaux objectifs de l'Association Américaine d'Évaluation incluent d'accroître l'utilisation de l'évaluation et de promouvoir l'évaluation en tant que profession. Pour promouvoir l'évaluation en tant que profession, il me semble que nous avons besoin d'une vision de notre profession comme étant source de changement, non pas dans un futur vague,

éloigné du domaine de l'évaluation, mais dans le sens plus immédiat de la réalisation des utilisations escomptées (qui sont les objectifs de l'évaluation) afin que les gens en aient pour leur argent.

J'ai pour l'évaluation la vision d'une profession qui s'acquitte de sa promesse de faire advenir les utilisations escomptées par les utilisateurs et utilisatrices ciblé-e-s – et qui documente par le biais d'études de suivi, les changements suscités par l'évaluation. Une telle profession aurait suffisamment confiance en son produit et ses processus, pour « vendre » l'évaluation (ou du moins la promouvoir avec enthousiasme). Une telle profession serait source de changement démontrable sur le plan de la qualité des programmes évalués et, par conséquent, de la qualité de vie des bénéficiaires de l'action publique.

Je crois qu'une telle profession existe aujourd'hui. Ma vision n'est pas celle d'un avenir lointain, mais une description de ce qui, selon moi, existe actuellement dans la plupart des pratiques d'évaluation. C'est pourquoi je vous ai invité-e-s aujourd'hui à vous joindre à moi pour célébrer l'évaluation.

Bibliographie

Etheredge, Lloyd S. 1979. « Government learning: an overview ». in *Handbook of Political Behavior*. New York: Plenum, p. 73-161.

Patton, Michael Q. 1978. *Utilization-Focused Evaluation*. 1re éd. Beverly Hills: Sage Publications.

Patton, Michael Q. 1986. *Utilization-Focused Evaluation*. 2e éd. Beverly Hills: Sage Publications.

Steele, Fritz. 1975. *Consulting for Organizational Change*. Amherst: University of Massachusetts Press.

Ziglar, Zig. 1987. « Closing skills ». *Podcast Personal Selling Power*, Episode 6.

3. Et si les responsables publics ne décidaient qu'en s'appuyant sur l'information : réponse à Patton

CAROL H. WEISS

[Traduit de : Weiss, Carol H. 1988. « If Program Decisions Hinged Only on Information: A response to Patton ». *American Journal of Evaluation*, 9(3) : 15-28 (Extraits). Traduction de Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Un bilan « neutre »

J'ai dit que nous avons eu un bilan « neutre » dans l'entreprise de « faire de l'évaluation la base de la décision ». Mike (Michael Q. Patton) affirme qu'il peut « sans difficulté nommer au moins 25, probablement 50 praticiens et praticiennes de l'évaluation » qui « ont un impact considérable sur les programmes et les politiques ». Je suis ravie que Mike puisse penser à 25 voire 50 évaluateurs et évaluatrices (sur combien? 100? 200?) qui ont un impact. Leviton et Boruch (1984) montrent que de nombreuses évaluations conduisent à des modifications de programmes. Mais ces déclarations ne réfutent pas mon point de vue. Les responsables des programmes et des politiques publiques n'utilisent pas systématiquement l'évaluation « comme base de décision ». Notez que je n'ai pas dit que le bilan de l'évaluation était mauvais; j'ai dit « neutre », ce qui, selon le Dictionnaire Webster signifie : « ni bon ni mauvais, ni grand ni petit, ni souhaitable ni indésirable, etc. ». Ce que je voulais dire par là, c'est

que même lorsque les évaluateurs et évaluatrices essaient d'adopter une démarche d'évaluation favorisant l'utilisation, leur influence reste sur le point médian d'une l'échelle de notation.

Je pense que les évaluateurs et évaluatrices peuvent faire mieux que ce qu'ils et elles font actuellement, pour attirer l'attention sur leurs résultats. Je pense même qu'ils et elles devraient le faire. Mais, dans l'ensemble, je doute que nous puissions un jour persuader les parties prenantes de faire des résultats de l'évaluation le paramètre essentiel des décisions relatives aux programmes publics. D'une part, les responsables des programmes en savent beaucoup plus sur leurs programmes que ce que leur dit l'équipe d'évaluation. Les responsables des programmes ont une expérience directe de l'organisation opérationnelle; ils et elles connaissent le site, les publics, le personnel, les problèmes, les budgets, les directives contradictoires des commanditaires et des bailleurs de fonds, l'état des relations avec d'autres organisations qui orientent ou accueillent les publics, l'histoire, les récriminations et les félicitations, ainsi que les perspectives d'avenir. Les équipes d'évaluation peuvent leur dire beaucoup de choses, mais, comme l'écrivent Gilsinan et Volpe (1986 : 182) : « Le chercheur ou la chercheuse en évaluation n'a le plus souvent qu'une occasion de concevoir et de mettre en œuvre … une étude [sur un programme donné], et dans des conditions suboptimales ». L'étude d'évaluation ne peut pas couvrir tous les aspects d'un programme et elle ne peut jamais être la seule base sur laquelle se fondent les décisions. Demander aux responsables de programmes et opérateurs et opératrices d'adopter pleinement les conclusions de l'évaluation revient à leur demander de mettre entre parenthèses leurs années d'expérience et d'immersion directe dans le monde quotidien du programme et, en fait, d'abdiquer leur responsabilité en faveur d'une équipe d'évaluation qui n'a inévitablement qu'une vision partielle de leurs préoccupations. Je doute que Mike ait eu l'intention de suggérer une telle démarche. Il contestait plutôt l'affect et la charge émotionnelle de ma déclaration, plutôt que mes mots au sens littéral (« faire de l'évaluation la base des décisions »).

Ce que les évaluateurs et évaluatrices devraient aspirer à atteindre dans le domaine de l'utilisation, c'est l'influence, et non le statut de philosophes-rois dont les diktats détermineraient l'avenir des programmes. Il est présomptueux de penser qu'une évaluation, aussi consciencieuse soit-elle, pourrait ou devrait constituer la base principale des changements apportés à un programme. (Je ne mentionnerai même pas les études d'évaluation effectuées à des niveaux de compétence médiocres ou pire.). Il est préférable de concevoir l'évaluation comme permettant de mieux comprendre le programme, de mettre en exergue l'éventail des options et des effets probables. En substance, l'évaluation devrait faire fonction de formation continue pour les responsables et opérateurs/-trices de programmes ainsi que pour les responsables politiques.

Le caractère omniprésent de la variable politique

Même lorsque nous parlons d'influence sur la conception et la mise en œuvre des programmes, il existe d'importantes raisons pour lesquelles les responsables de programmes ne prêtent pas toujours une grande attention aux résultats de l'évaluation. Par la conduite de leur programme, ils et elles poursuivent bien d'autres intérêts, en outre de l'exécution de la mise en œuvre planifiée et de l'obtention des résultats escomptés. Ils et elles veulent avoir une vie professionnelle satisfaisante, s'entendre avec leurs collègues, être reconnu-e-s et respecté-e-s et voir leur organisation gagner en prestige et en solvabilité financière, peut-être surpasser une agence ou une faction concurrente, avoir une possibilité d'ascension, faire un travail qui soit estimé par les membres de leur profession, respecter des traditions et s'amuser. Lorsque les résultats et les recommandations des évaluateurs et évaluatrices mettent en péril de telles valeurs, les responsables de programmes placent parfois leurs propres intérêts au premier plan.

L'intérêt personnel, la protection de l'organisation, la recherche de bénéfices : ces éléments sont remarquablement absents du monde de Patton. Dans un manuscrit sur l'utilisation des évaluations qui compte 25 pages dactylographiées, Mike Patton ne mentionne jamais le mot « politique ». Dans son monde, tout le monde se comporte rationnellement. Non seulement les responsables des programmes et des politiques publiques sont tou-te-s rationnel-le-s, prêt-e-s à fonder leurs décisions sur les meilleures données disponibles si l'équipe d'évaluation est suffisamment persuasive et persistante, mais ils et elles ont aussi des motivations altruistes. Ils et elles veulent, par-dessus tout, améliorer leur programme pour servir les intérêts des bénéficiaires et ne semblent pas se préoccuper des amputations budgétaires, de la recherche de personnel qualifié, du travail supplémentaire, de la perturbation des relations en cours avec d'autres organismes, des éventuelles réactions négatives des groupes communautaires ou de la presse, du renouvellement de leur subvention, de la satisfaction des rancunier-e-s du conseil d'administration ou de toute autre préoccupation qui tourmente les responsables de programmes avec lesquel-le-s j'ai eu affaire. Ses parties prenantes sont prêtes à utiliser les évaluations si l'équipe d'évaluation présente un dossier suffisamment solide. Ils ne s'inquiètent pas de la rareté des ressources, de la réputation de leurs programmes, de leur propre avancement ou de l'évitement de tâches désagréables. Ses responsables politiques ne s'inquiètent pas de la prochaine élection, de l'obtention de crédits plus importants, ou de ménager les susceptibilités des législateurs/-trices ou des administrateurs/-trices influent-e-s.

Selon Patton, tout ce que l'équipe d'évaluation a à faire est de leur communiquer les faits et de leur indiquer la « bonne » ligne de conduite. Il est vrai que l'évaluatrice ou l'évaluateur peut avoir à le dire à maintes reprises et de différentes manières, et qu'il ou elle doit être un-e vendeur/-euse, un-e charmeur/-euse, une personne dotée d'excellentes compétences interpersonnelles. L'équipe d'évaluation doit croire en l'efficacité du produit qu'elle vend, dit Mike, et apprendre les techniques de vente. Toutefois, en accordant suffisamment d'attention aux premiers

contacts, à l'implication du ou de la praticien-ne, à la participation du commanditaire à l'évaluation et à une bonne diffusion, l'équipe d'évaluation aura un impact considérable sur les décisions des responsables et des décideurs/-euses rationnel-le-s et intelligent-e-s.

Permettez-moi de vous parler de certaines évaluations auxquelles j'ai participé.

(1) J'ai dirigé une évaluation d'un programme financé par le gouvernement fédéral au sein d'un organisme d'action sociale. Peu après la publication de l'étude d'évaluation, la subvention fédérale a pris fin. Aucun autre organisme local n'avait les ressources nécessaires pour reprendre le programme (même s'il avait été très efficace). Le personnel de l'unité de financement de Washington qui aurait dû être intéressé par les résultats a été submergé de rapports et s'inquiétait de son propre avenir. Utilisation : nulle.

(2) Une évaluation d'un million de dollars à laquelle j'ai participé à titre de consultante a fourni une quantité considérable d'informations utiles sur le processus et les résultats du programme. Elle a montré que le programme avait un succès modeste dans l'amélioration de la pratique médicale dans certains domaines, un succès moindre dans d'autres domaines, et elle a indiqué les stratégies qui affectaient l'efficacité du programme. Au moment où le travail s'achevait, une nouvelle personne a été nommée à la tête de l'agence mère, et ses priorités n'incluaient pas le programme étudié. Le directeur du programme a été encouragé à partir, et le programme a été relégué à un statut périphérique au sein de l'agence. Toutes les preuves de succès et les recommandations importantes sur les orientations à prendre pour améliorer le programme ont trouvé leur place dans les publications professionnelles, d'où elles pourront peut-être resurgir un jour pour influer sur la prochaine réincarnation du programme.

(3) Une étude d'évaluation d'un programme dans un petit organisme a été entreprise parce qu'il y avait une forte divergence d'opinions parmi le personnel sur l'intérêt de consacrer des ressources importantes à ce programme au détriment d'autres que l'organisme administrait ou souhaitait mettre en œuvre. Lorsque l'évaluation a été publiée, les partisan-e-s du programme se sont emparé-e-s des résultats positifs et les opposant-e-s des résultats négatifs. Le débat au sein de l'agence a continué à faire rage, mais chaque partie citait maintenant des preuves d'évaluation pour justifier son argumentation.

(4) Une autre étude d'évaluation a révélé qu'une modalité de mise en œuvre du programme avait des effets nettement supérieurs aux autres modes, et l'évaluateur a recommandé l'expansion des stratégies efficaces. Toutefois, ce type de programmation nécessitait également beaucoup plus de personnel, et les coûts de fonctionnement étaient près d'un tiers plus élevés que les coûts habituels. L'expansion de la stratégie la plus efficace impliquait également de réduire le nombre de personnes bénéficiaires, avec toute la publicité négative qu'une telle décision entraînerait. L'agence a donc décidé de procéder comme auparavant, en promettant de faire des ajustements mineurs vers le meilleur procédé à mesure que le budget le permettrait.

Tous les évaluateurs et évaluatrices que je connais ont vécu des expériences similaires. Pas à chaque fois, bien sûr, sinon ils et elles auraient tous fui le terrain ou seraient devenus des cyniques confirmé-e-s, mais assez souvent pour reconnaître le scénario. La doctrine « Patton » de la responsabilité voudrait que les évaluateurs et évaluatrices soient responsables de ces échecs. D'une manière ou d'une autre, s'ils et elles avaient été véritablement à la hauteur, ils et elles auraient pu prévoir ou atténuer les pressions hostiles. J'aimerais bien savoir comment.

[…]

Fiabilité des preuves

Sur quoi Mike fonde-t-il sa conviction que le niveau d'utilisation des évaluations est élevé? Principalement sur la base de comptes rendus des équipes d'évaluation sur leur propre succès. En tant qu'évaluateurs et évaluatrices, nous avons tous eu connaissance de programmes dans lesquels le personnel nous dit à quel point ce qu'ils et elles font est un succès. Nous restons sceptiques face à ces témoignages et les soumettons à un test empirique. Comme me l'a dit il y a longtemps le directeur d'un programme de santé mentale : « le travail d'un-e praticien-ne de programme est de croire; le travail d'un évaluateur ou d'une évaluatrice est de douter ». En tant que praticien-ne-s de l'évaluation, nous avons tendance à croire en la valeur de l'entreprise et à minimiser les cas d'échec. Mais en tant qu'évaluateurs et évaluatrices, nous sommes obligé-e-s d'examiner les preuves.

Les preuves de Mike proviennent, en grande partie, d'articles que les évaluateurs et évaluatrices ont écrits à l'intention de publics professionnels, sur leurs succès en matière d'utilisation. La plupart d'entre nous peuvent aussi écrire de tels articles, si nous voulons montrer le côté positif de l'histoire. C'est une partie de la vérité, mais en aucun cas toute la vérité.

Un autre type de preuve que Mike cite est le suivi par les unités d'évaluation du sort de leurs recommandations. Par exemple, les unités du *General Accounting Office* (GAO) et du *Federal Bureau of Investigation* (FBI) ont compté le nombre de leurs recommandations qui ont été acceptées et mises en œuvre, et les taux sont élevés. Ayant moi-même participé à un exercice de ce genre, je considère ces données avec respect, mais aussi avec un certain scepticisme. Dans un cas, je me souviens, nous avons suivi les utilisations faites d'une étude qui avait formulé cinq recommandations. L'une d'entre elles portait sur une refonte majeure du programme, et les quatre autres concernaient l'amélioration de la tenue

des dossiers, des procédures budgétaires, des pratiques comptables et des rapports. L'organisme a mis en œuvre quatre des cinq recommandations (devinez lesquelles). Son taux d'utilisation a été de 80%!

De toute façon, je suis toujours un peu mal à l'aise à l'idée de prendre les recommandations comme unité d'utilisation. À ma connaissance, nous n'avons pas examiné de très près l'origine des recommandations des équipes d'évaluation. Certaines recommandations peuvent être bien fondées sur des données, tandis que d'autres peuvent être des envolées fantaisistes de la part de personnes n'ayant pas beaucoup d'expertise dans la planification ou le fonctionnement des programmes. Permettez-moi de vous faire part de certaines de mes réflexions sur les sources des recommandations. C'est un pas de côté par rapport au thème principal de cet article, mais cela pourrait inciter certain-e-s d'entre vous à étudier attentivement le sujet.

Dans les meilleurs cas, les recommandations découlent directement des données. Il y a des preuves tangibles qu'une pratique du programme est meilleure que d'autres pratiques du programme, et l'évaluateur ou l'évaluatrice recommande celle-ci. Par exemple, si les étudiant-e-s qui consacrent plus de temps à étudier ont tendance à apprendre davantage, la recommandation de passer plus de temps à cette tâche est bien fondée. La plupart du temps, j'imagine, les recommandations représentent un bond en avant par rapport aux données. Elles peuvent découler de normes ou de lignes directrices professionnelles. Par exemple, lorsqu'un programme n'est pas particulièrement efficace et que l'équipe d'évaluation sait qu'il ne respecte pas la « bonne pratique » dans le domaine ou dans la profession, l'équipe d'évaluation recommande une bonne pratique. Parfois, les recommandations semblent être faites parce que ce que fait le programme ne fonctionne pas très bien, et l'équipe d'évaluation suppose qu'il serait préférable de faire le contraire. Par exemple, si le programme infructueux utilise le conseil collectif, l'équipe d'évaluation peut recommander le conseil individuel; si le programme

infructueux repose sur des incitations pour les travailleurs et travailleuses individuel-le-s, l'équipe d'évaluation peut recommander des incitations collectives.

Dans certains cas, les évaluateurs et évaluatrices possèdent des connaissances spécialisées dans le domaine du programme. Ils et elles ont étudié de nombreux programmes d'éducation compensatoire, de réadaptation physique ou de placement en famille d'accueil, et (avec ou sans formation professionnelle) ont développé un répertoire de connaissances. Leurs recommandations découlent de cet ensemble de compétences. Le plus souvent, je pense que les évaluateurs et évaluatrices s'appuient sur un raisonnement logique. Ils et elles essaient de comprendre ce qu'il faudrait faire pour améliorer le fonctionnement d'un programme. Ils et elles élaborent implicitement un modèle logique du programme dans leur esprit, corrigeant les lacunes et les incohérences qu'ils et elles voient dans le programme, et fondent leurs recommandations sur leur compréhension de profane.

Il n'est pas rare que l'équipe d'évaluation consacre tellement de temps à la collecte et à l'analyse de données qu'il lui reste très peu de temps à la fin pour comprendre les implications de celles-ci. À deux semaines de la remise du rapport, elle se démènent pour trouver quelque chose de raisonnable à recommander. La pertinence de leurs recommandations dépendra fortement de leur niveau d'information sur le terrain, sur d'autres programmes et évaluations antérieures, ainsi que sur les comportements individuels et collectifs. Pour les personnes participant au programme, qui sont les utilisateurs et utilisatrices visé-e-s par les conclusions et les recommandations, prendre les recommandations de l'équipe d'évaluation au sérieux peut être soit un coup brillant, soit un exercice massif de futilité. Je suis plus à l'aise avec l'idée qu'ils et elles prennent les résultats de l'évaluation au sérieux. Les responsables de programmes et les praticien-ne-s savent peut-être mieux que l'équipe d'évaluation quelles sont les conséquences à tirer des preuves et quelles directions sont susceptibles d'être les plus fructueuses – ce qui est un bon point pour revenir au sujet principal de cet article.

Dans l'ensemble, je pense que le degré d'utilisation directe des résultats d'évaluation par les organisations est – si vous n'aimez pas le mot « neutre » – pas mauvais, honnête, une chose utile de temps à autre. Si les résultats leur montrent qu'il y a quelque chose à corriger et une façon de le faire, elles essaieront souvent de le faire. S'ils fournissent une pièce supplémentaire du puzzle de l'action de l'organisation, elle prendra sa place dans l'ensemble. S'ils ne correspondent pas à ce que les organisations croient et savent, ou croient savoir, elles peuvent, après réflexion, les classer pour plus tard. [...]

Bibliographie

Gilsinan, James F., et L. Carl Volpe. 1986. « Do not cry wolf until you are sure ». *Evaluation Studies Review Annual* 11 : 175-87.

Leviton, Laura C., et Robert F. Boruch. 1984. « Contributions of evaluation to education programs and policy ». *Evaluation Studies Review Annual* 9 : 597-632.

4. Repenser l'utilisation de l'évaluation. Une théorie intégrée de l'influence

KAREN E. KIRKHART

[Traduit de : Kirkhart, Karen E. 2000. « Reconceptualizing evaluation use: An integrated theory of influence ». *New Directions for Evaluation*, 2000(88) : 5–23 (Extraits). Traduction par Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de John Wiley and Sons.]

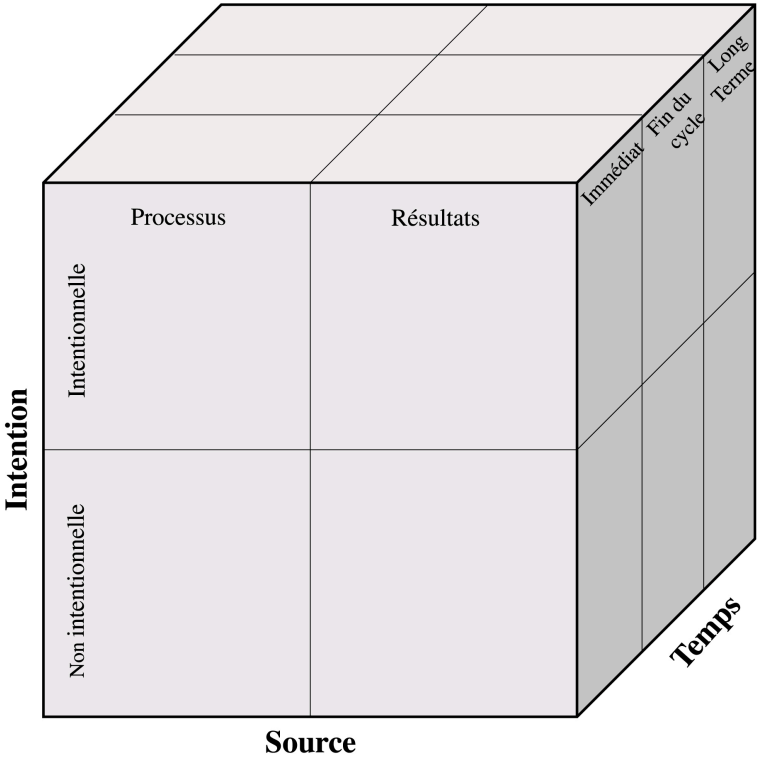
Une théorie intégrée de l'influence

Une théorie intégrée de l'influence comprend trois dimensions : la source d'influence, l'intention et le temps. Chaque dimension est subdivisée en niveaux (voir Figure 1). Ces subdivisions sont certes quelque peu arbitraires. La source, l'intention et le temps peuvent être caractérisés de façon plus précise comme un *continuum*, reflétant les zones grises qui se situent entre les niveaux.

La source d'influence désigne l'agent actif du changement ou le point de départ d'un processus générant du changement (Henry et Rog, 1998). La source d'influence se distingue en deux niveaux, les influences associées au processus d'évaluation et des influences associées aux résultats de l'évaluation. L'intention se rapporte à la mesure dans laquelle il existe une orientation intentionnelle visant à exercer un type particulier d'influence dans le cadre du processus d'évaluation ou des résultats d'évaluation. Elle reflète l'importance de tenir compte à la fois de l'influence escomptée et non intentionnelle de l'évaluation. La division de la troisième dimension,

celle du temps, en influence immédiate, en fin de cycle et à long terme reflète la nécessité de reconnaître l'influence pendant et immédiatement après le cycle d'évaluation ainsi que les effets visibles à l'avenir. Ces trois dimensions et leurs antécédents historiques sont résumés dans les sections suivantes.

Figure 1 : Théorie intégrée de l'influence



Source d'influence

La première dimension porte sur la dimension de l'évaluation qui est présumée exercer un pouvoir ou une influence sur des individus, des organisations ou des collectivités décisionnelles plus larges. Historiquement, l'influence de l'évaluation était définie en fonction de l'utilisation des résultats (Johnson, 1998; Shulha et Cousins, 1997). Il y avait une littérature parallèle qui s'intéressait à l'influence interpersonnelle du processus d'évaluation, mais ces deux volets n'étaient pas réunis avant la « découverte » de l'utilisation du processus (Patton, 1998). L'utilisation du processus d'évaluation est apparue d'abord comme un moyen de faciliter l'utilisation axée sur les résultats, puis a été considérée comme une source d'influence en soi. Toutefois, les vestiges de cette évolution subsistent, sont reflétés dans le langage d'utilisation, et créent des confusions conceptuelles. L'utilisation du processus est souvent liée à tort à la typologie reconnue de l'utilisation axée sur les résultats plutôt que d'être considérée comme une autre typologie à part entière. Le passage de l'utilisation à l'influence crée un cadre pour un traitement parallèle des deux dimensions, reflété dans ce modèle.

L'influence basée sur les résultats. L'attention accordée dès le début à l'utilisation de l'évaluation découlait d'un désir de maximiser l'impact social positif de l'évaluation, en lien avec des préoccupations quant à la non-utilisation des résultats de l'évaluation (Ciarlo, 1981). Que la question soit examinée du point de vue des acteurs et actrices impliqué-e-s dans l'élaboration des politiques publiques (Cronbach, 1982; Cronbach *et al.*, 1980), des utilisateurs et utilisatrices individuel-le-s, des décisions prises, de l'organisation à évaluer ou des rapports publiés (Weiss 1981), les points de référence pour juger de l'utilisation de l'évaluation étaient l'information produite par l'évaluation et les conclusions fondées sur des données. Dans leur synthèse des travaux empiriques sur l'utilisation de l'évaluation, Cousins et Leithwood (1986) définissent les résultats de l'évaluation comme « toute information liée aux résultats de l'évaluation; par exemple, données, interprétations, recommandations; ces informations pourraient

être communiquées à l'issue de l'évaluation ou au fur et à mesure que celle-ci se poursuit » (p. 332). Bien que cette définition distingue clairement des utilisations formative ou sommative, les premiers travaux empiriques se sont concentrés sur l'utilisation sommative. L'utilisation basée sur les résultats a d'abord été considérée sous l'angle de l'utilisation instrumentale, c'est-à-dire à partir des mesures directes et visibles adoptées sur la base des résultats de l'évaluation (Rich, 1977). Cette conceptualisation étroite s'est rapidement élargie pour inclure les utilisations conceptuelles des résultats, comme l'apport d'un éclairage nouveau (illumination) et la démystification. A ainsi été intégré un impact cognitif sur les appréciations ou les compréhensions qui ne se traduisait pas nécessairement par un changement de comportement manifeste (Rich, 1977; Weiss et Bucuvalas, 1980). Un troisième type d'utilisation, axée sur les résultats, a porté sur le rôle des conclusions de l'évaluation dans le plaidoyer, l'argumentation et le débat politique (Greene, 1988a; Knorr, 1977; Leviton et Hughes, 1981; McClintock et Colosi, 1998; Johnson, 1998; Shadish, Cook, et Leviton, 1991). Que ce soit l'utilisation légitimatrice, l'usage symbolique, l'usage politique ou l'usage persuasif, cette application se concentrait explicitement sur l'utilisation de l'information évaluative pour convaincre les autres de promouvoir une position ou de défendre contre des attaques une option déjà adoptée. Ensemble, ces trois vecteurs d'utilisation ont délimité le paysage conceptuel de l'influence fondée sur les résultats.

L'influence basée sur le processus. L'influence de l'évaluation ne découle pas entièrement de la présentation formative ou sommative des résultats. Parfois, l'influence principale se joue dans le processus de réalisation de l'évaluation en tant que tel. Bien que le terme « processus » n'ait pas été utilisé dans la littérature sur l'évaluation avant la fin des années 1980, l'attention portée à l'impact du processus d'évaluation peut être identifiée dès les premiers travaux portant sur les rôles d'agent-e de changement de l'évaluateur ou de l'évaluatrice, dans la recherche-action et dans des approches telles que l'évaluation transactionnelle, qui ont mis l'accent

sur les interactions entre les évaluateurs/-trices et des systèmes ouverts (Anderson et Ball, 1978; Argyris, Putnam, et Smith, 1985; Caro, 1980; Rippey, 1973; Rodman et Kolodny, 1972; Tornatzky, 1979). Les racines et l'évolution de l'influence basée sur le processus sont particulièrement visibles dans l'histoire des approches d'évaluations participatives (Brisolara, 1998).

L'utilisation qui découle du processus a d'abord été théorisée, dans la littérature sur l'utilisation, comme un moyen de faciliter l'utilisation basée sur les résultats. L'objectif de Greene (Greene, 1988b) par exemple, était de créer des conditions propices à une utilisation basée sur les résultats plutôt que sur les effets intrinsèques du processus d'évaluation lui-même. Par la suite, les réflexions sur l'utilisation qui découle du processus ont mis l'accent sur sa valeur indépendamment de l'utilisation basée sur les résultats (Whitmore, 1991). Patton (1997) la décrit comme « des façons par lesquelles la participation au processus d'évaluation peut être utile indépendamment des résultats qui peuvent découler de ces processus » (p. 88, mention en italique de l'autrice).

Greene (1988b) distingue trois dimensions de l'influence basée sur le processus : cognitive, affective et politique. La dimension cognitive fait référence à des changements dans la compréhension des choses, stimulés par la discussion, la réflexion et l'analyse des problèmes enchâssées dans le processus d'évaluation. Bien que l'utilisation dérivée du processus puisse comporter une composante instrumentale (comme lorsque la réflexion mène à une décision ou à une action), la dimension cognitive met l'accent sur une meilleure compréhension de la politique publique parmi les participant-e-s au processus d'évaluation. La dimension affective est plus personnellement liée aux participant-e-s eux-mêmes. Cette dimension porte sur les sentiments individuels et collectifs de valeur qui découlent du processus d'évaluation. Bien que les interprétations de Greene (1988b) identifient des sentiments relatifs à l'estime de soi, d'autres mécanismes psychologiques sont également possibles (par exemple, les sentiments à l'égard de l'évaluation, les sentiments à l'égard du programme lui-même). La dimension politique

porte sur l'utilisation du processus d'évaluation lui-même pour générer de nouvelles discussions, attirer l'attention sur des problèmes sociaux ou influencer la dynamique du pouvoir et des privilèges inhérents à l'évaluation et à son contexte. Les modèles récents qui définissent l'évaluation comme une intervention visant explicitement à modifier la mise en œuvre ou à rendre compte des résultats d'une politique publique soulignent l'importance de la dimension politique de l'influence qui découle du processus (Fetterman, 1994; Patton, 1998). Ces trois types d'influence basée sur le processus peuvent opérer ensemble. La discussion de Shulha (dans ce volume[1]) sur ce que les évaluateurs et évaluatrices eux/elles-mêmes ont appris en participant au travail d'évaluation qu'elle comporte des aspects cognitifs, affectifs et politiques.

Résumé. La première dimension d'une théorie intégrée de l'influence, c'est-à-dire la source d'influence, s'intéresse à l'élément de l'évaluation qui est présumé générer des changements. Au sens large, les deux sources d'influence sont le processus d'évaluation et les résultats qui sont générés par l'évaluation. Des démarches d'évaluation particulières peuvent très bien recouvrir les deux types d'influence; toutefois, de nombreuses démarches accordent une importance différente aux deux sources d'influence, et certaines se concentrent presque exclusivement sur une seule source. Bien que les typologies d'utilisation se soient développées pour chaque source d'influence, ces distinctions ne devraient pas éclipser les liens entre l'influence basée sur le processus et l'influence basée sur les résultats. Non seulement les deux types d'influence peuvent être combinés dans un modèle logique unique (voir, par exemple, Greene, 1988b), mais les sous-catégories de chaque source d'influence peuvent être interdépendantes. Anderson, Ciarlo et Brodie (1981), par exemple, discutent des dimensions affectives de l'utilisation

1. Special Issue: *The Expanding Scope of Evaluation Use. New Directions for Evaluation*, Volume 2000, Issue 88, Winter 2000.

basée sur les résultats, tandis que Greene (1988b) fait allusion à l'impact instrumental de l'utilisation du processus. Dans ce volume, la structure de l'argument de Henry illustre un lien implicite entre l'utilisation basée sur les résultats et l'utilisation basée sur le processus. L'objet principal de son chapitre est l'influence basée sur les résultats, mais il est intéressant de noter que selon lui seule une utilisation dérivée du processus de l'évaluation peut amener à la mise à l'agenda de nouveaux problèmes publics.

L'intention

L'intention est la deuxième dimension d'une théorie intégrée de l'influence de l'évaluation (Kirkhart, 1999). Elle désigne dans quelle mesure l'influence de l'évaluation est intentionnellement dirigée, reconnue consciemment et planifiée. Les plus visibles sont les influences prévues qui sont explicites dans le but de l'évaluation, dans la théorie utilisée et dans le contrat évaluateur-commanditaire. Les objectifs latents et les agendas d'évaluation cachés peuvent aussi relever de l'intention, mais ces intentions peuvent être plus difficiles à cerner. Les influences imprévues s'intéressent aux répercussions non intentionnelles de l'évaluation sur les individus et les systèmes, souvent par des voies inattendues. Une évaluation donnée peut n'avoir qu'une influence intentionnelle, une influence non intentionnelle ou un mélange des deux. La cartographie des influences tant prévues qu'inattendues est essentielle à l'appréciation complète de l'impact de l'évaluation.

L'intention d'influencer (diversement appelée intention et intentionnalité) a joué un rôle important dans la conceptualisation de l'utilisation de l'évaluation. C'est l'une des premières dimensions identifiées dans l'élaboration d'une théorie intégrée de l'influence (Kirkhart, 1995). Selon d'autres théories d'utilisation, elle marque la frontière entre l'utilisation et les mésusages (*misuse*) de l'évaluation (Alkin, 1990; Alkin, Daillak, et White,

1979; Christie et Alkin, 1999). La question qui définit cette deuxième dimension de l'influence est : Quelles sont les intentions de l'équipe d'évaluation, du/de la commanditaire et d'autres intervenant-e-s clés concernant l'influence de l'évaluation? L'intention peut être encore décomposée en trois aspects : le type d'influence souhaité ou prévu; qui doit être influencé; ainsi que les personnes, le processus et les résultats qui sont censés exercer une influence. Bien que les deux premiers soient souvent confondus, une réflexion distinguant les deux facilite l'identification des influences involontaires.

L'influence intentionnelle. L'évaluation peut avoir pour but d'exercer une influence soit par le processus lui-même, soit par les résultats obtenus. La notion « d'utilisation primaire » de Patton (1997), *utilisation escomptée par les utilisateurs ciblés*, marque un chemin direct entre l'intention et l'influence. Historiquement, le scénario le plus souvent envisagé est basé sur l'utilisation des résultats. Les utilisateurs et utilisatrices potentiel-le-s des connaissances produites par l'évaluation sont identifié-e-s dès le début du processus d'évaluation, et leurs besoins de connaissance déterminent l'évaluation, depuis les questions posées jusqu'aux données recueillies et à la façon dont les résultats sont communiqués. Toutefois, il peut y avoir une intention tout aussi explicite d'influencer les organisations et les systèmes sociaux par le biais du processus d'évaluation lui-même, comme l'illustre l'évaluation participative. Cousins et Whitmore (1998) distinguent deux courants d'influence basés sur le processus dans l'évaluation participative, chacun ayant sa propre idéologie et sa propre intention. Dans le cadre de l'évaluation participative transformatrice, l'objectif est l'autonomisation, l'action sociale et le changement, tandis que l'évaluation participative pratique vise à aider à la résolution de problèmes au niveau des programmes ou de l'organisation. De même, Patton (1998) décrit « l'évaluation comme une intervention intentionnelle en vue d'améliorer les résultats du programme » (p. 229),

qui est une sorte d'utilisation de processus, conceptualisant la réactivité naturelle du processus de collecte de données comme une intervention qui renforce ce que le programme essaie de faire.

Une mise en garde importante concernant l'influence visée est que toutes les intentions ne sont pas explicitement communiquées ou rendues visibles. Les objectifs déclarés d'une évaluation correspondent à ses fonctions manifestes, évidentes; par exemple, une évaluation formative peut être entreprise dans le but d'améliorer l'intervention publique, tandis qu'une évaluation sommative peut être entreprise pour aider un promoteur à mieux réaffecter les fonds. Toutefois, les influences prévues pourraient aussi inclure des fonctions cachées et latentes (Scriven, 1991). Par exemple, une évaluation ayant pour fonction manifeste d'améliorer l'efficacité du programme pourrait aussi avoir pour fonction latente d'accroître la visibilité du programme auprès de la collectivité. Une évaluation ayant manifestement pour fonction de démontrer la responsabilité et l'efficacité des commanditaires pourrait entraîner une réaffectation des fonds et une réduction des effectifs en tant qu'objectifs latents. Pour saisir toute la gamme d'influences prévues, il faut prêter attention aux fonctions manifestes et latentes. Ici, la pluralité des utilisations prévues et des utilisateurs et utilisatrices devient critique. Il faut tenir compte de la compréhension et des agendas des commanditaires de l'évaluation, des évaluateurs/-trices et des parties prenantes.

L'influence non intentionnelle. L'évaluation peut influencer les programmes et les systèmes d'une manière non intentionnelle, par des voies imprévues. L'attention portée à l'influence involontaire de l'évaluation reconnaît à la fois le pouvoir des effets d'entraînement et notre incapacité à anticiper toutes les ramifications de notre travail. Dans le cadre de l'évaluation, comme dans les programmes eux-mêmes, l'influence non intentionnelle peut avoir plus d'impact que l'influence prévue. En outre, le territoire défini par une influence involontaire est

plus large. Alors que l'utilisation primaire attire plutôt l'attention sur les utilisations prévues et les utilisateurs, l'influence involontaire prend la forme d'une série de permutations.

Trois variantes illustrent ce point. Premièrement, les utilisateurs et utilisatrices ciblé-e-s peuvent exercer une influence de manière non intentionnelle ou affecter des personnes ou des systèmes autres que ceux prévus. Prenons un exemple d'utilisation basé sur les résultats. Un groupe consultatif est l'un des utilisateurs ciblés par les conclusions de l'évaluation. Bien que l'utilisation attendue de leur part consistât à apporter des changements internes au programme, les données ont eu des répercussions inattendues sur le personnel politique qui a formé une coalition communautaire pour plaider en faveur d'un changement législatif. Cette influence plus large n'était pas intentionnelle, bien qu'elle ait été initiée par les utilisateurs et utilisatrices ciblé-e-s. Deuxièmement, des utilisateurs et utilisatrices non prévu-e-s peuvent être impliqué-e-s dans l'exercice de l'influence, bien que la nature de l'influence et que d'autres personnes et systèmes concernés soient ciblés. Prenons un exemple de processus dans lequel une évaluation des besoins est effectuée sur le problème de la violence dans les écoles publiques. L'intention était de faire participer les parents et les enseignant-e-s avec les membres des conseils d'écoles à l'identification des préoccupations et à la formulation de solutions en vue d'un environnement scolaire sécurisé; toutefois, les élèves ont affirmé leur intérêt pour l'évaluation, et leur participation à l'évaluation des besoins a modifié le climat de l'école. L'influence va dans le sens intentionnel (vers la sécurité) et sur le système prévu, mais elle est passée par un chemin d'utilisation non prévu. Troisièmement, les utilisateurs et utilisatrices, la nature de l'influence et les systèmes influencés peuvent tous être involontaires. Prenons par exemple l'évaluation interne d'un organisme local d'accompagnement des publics visant à appuyer une demande de prolongation du financement par l'organisation communautaire qui le soutient. Au fil de l'évaluation, les bénéficiaires ont joué un rôle inattendu dans le processus, ce qui a suscité une publicité positive involontaire pour l'organisme. Le processus

d'évaluation inclusive a été cité comme modèle, et un groupe de défense des bénéficiaires à l'échelle de l'État a mis les promoteurs publics au défi de repenser les paramètres de l'évaluation dont ils avaient besoin pour obtenir un financement. Il est à noter que cette influence non intentionnelle peut s'ajouter à l'utilisation prévue des données pour appuyer la pérennité du financement.

Résumé. L'intention est la deuxième dimension clé d'une théorie intégrée de l'influence exercée par l'évaluation. Les influences prévues peuvent être basées sur les résultats ou sur le processus, manifeste ou latent. Les influences non intentionnelles peuvent aussi être liées au processus ou aux résultats; cependant, la nature de l'influence, les personnes ou les systèmes influencés et les personnes qui exercent l'influence sont autres que celles escomptées ou prévues. Les influences intentionnelles et non intentionnelles peuvent se produire individuellement ou se combiner et, comme l'indiquent les exemples cités précédemment, elles peuvent s'exercer à différents moments. Bien que les exemples présentés illustrent des influences positives, l'intention ne présage pas de la valeur positive de l'influence. De toute évidence, l'évaluation peut avoir des influences négatives involontaires sur les personnes ou les systèmes; certaines des influences prévues peuvent aussi avoir des répercussions négatives sur certaines parties du système. Prises ensemble, les trois dimensions de l'influence offrent un cadre permettant d'examiner à la fois les effets positifs et négatifs de l'évaluation.

Le temps

La troisième dimension, le temps, se rapporte à la chronologie cours de laquelle l'influence de l'évaluation apparaît, existe, ou se poursuit. Cette dimension met en évidence la nature dynamique de l'influence et la possibilité que différentes dimensions de l'influence se produisent à

différents moments, dans l'immédiat, à la fin du cycle de l'évaluation et à long terme. Étant donné que le temps est un *continuum*, cette subdivision en trois périodes est arbitraire, mais les catégories attirent l'attention sur l'influence à trois étapes différentes qui vont de pair avec l'opinion selon laquelle les résultats du programme sont immédiats, apparaissent à la fin de l'intervention ou à long terme (Scriven, 1991). Tout comme ces distinctions ont été utiles pour orienter l'attention des évaluateurs sur les résultats du programme, une convention semblable peut guider la conceptualisation de l'influence de l'évaluation.

Comme les deux dimensions précédentes, il y a des antécédents au traitement actuel du temps dans la littérature sur l'évaluation. Shadish, Cook et Leviton (1991) décrivent l'historique de l'utilisation comme étant passé d'une utilisation à court terme à la reconnaissance de l'utilisation à long terme. Cette dichotomie a été couramment utilisée pour décrire la dimension temporelle. Parmi les six dimensions clés qu'elle identifie pour conceptualiser l'utilisation de l'évaluation, Weiss (1981) inclut : « *Dans quelle mesure l'utilisation est-elle immédiate*? » (en opposant l'utilisation immédiate et l'utilisation à long terme) était l'une des six dimensions clés de la conceptualisation de l'utilisation. De même, Smith (1988) mentionne l'utilisation immédiate ou à long terme comme l'une des quatre dimensions caractérisant l'utilisation, bien que son analogie comparant l'utilisation de l'évaluation à l'emprunt de livres dans une bibliothèque repose exclusivement sur l'utilisation axée sur les résultats. Tou-te-s les auteurs et autrices n'ont pas traité le temps comme une dichotomie entre le court terme et le long terme. Wollenberg (1986, cité dans Johnson, 1998), dans une étude sur l'utilisation qui s'étendait sur une année scolaire complète, a conceptualisé la dimension temporelle selon trois périodes ou cycles caractérisant la mise en œuvre ou le développement du programme – stade conceptuel, stade de développement et stade institutionnel. Cronbach (1982) a décrit quatre périodes d'influence, faisant remarquer : « Une évaluation alimente la pensée sociale telle

qu'elle est planifiée, puisqu'elle apporte des données, au moment où elle touche à sa fin, et, on peut l'espérer, pendant plusieurs années par la suite » (p. 318).

Historiquement, les trois dimensions d'une théorie intégrée de l'influence – la source, l'intention et le temps – sont liées. La dimension temporelle a été attachée à d'autres distinctions d'utilisation d'une manière qui a brouillé toute la gamme d'influences chronologiques. La recension de Leviton et Hughes (1981) illustre comment la dimension temporelle a souvent été intégrée dans les premières discussions sur l'utilisation instrumentale par rapport à l'utilisation conceptuelle basée sur les résultats. Ils citent Rein et White (1975) pour marquer la reconnaissance précoce du fait que « les problèmes du gouvernement sont définis graduellement au fil du temps et que les décisions sont finalement prises sur la base d'un ensemble intégré d'information provenant de nombreuses sources » (Leviton et Hughes, 1981 : 531). Le fait que cette citation soit mobilisée pour illustrer l'utilisation instrumentale par rapport à l'utilisation conceptuelle témoigne de la confusion historique des dimensions – en l'occurrence, le temps avec la source. De même, dans leur recension des théories d'utilisation des principaux théoriciens de l'évaluation, Shadish, Cook et Leviton (1991) se réfèrent à maintes reprises à l'utilisation instrumentale à court terme et à l'utilisation conceptuelle à long terme. Bien qu'il s'agisse là de combinaisons claires et peut-être communes d'utilisation dans un cadre d'utilisation basée sur les résultats, la dimension temporelle devrait être examinée séparément de la nature de l'influence pour une clarté maximale.

Un point clé qui est pertinent pour la conceptualisation de la dimension temporelle est de savoir si l'utilisation est considérée comme un événement ponctuel ou comme un processus plus ouvert. Les premières définitions ont parlé de l'utilisation comme d'un événement. Par exemple, Alkin, Daillak et White (1979) ont demandé : « Comment savoir identifier une utilisation? » et « Comment définissons-nous une utilisation? » (p. 226)? Leur modèle d'utilisation a culminé dans la description de « cas d'utilisations » (p. 232). Leviton et Hughes (1981) ont utilisé ce même

langage, en se demandant : « Qu'est-ce qu'un cas d'utilisation? » (p. 533), le cas correspondant à l'unité d'analyse dans leur méthode d'étude de l'utilisation. Ailleurs, une évolution se faisait jour, d'une conception de l'utilisation comme un événement distinct à son appréhension comme un processus ouvert. En remplaçant le terme utilisation par usages, Weiss (1981) a souligné la nécessité de s'éloigner d'une conception de l'utilisation comme événement, notant que l'utilisation de l'évaluation différait de celle d'un marteau. Cronbach (1982) considère l'utilisation comme un processus, et non comme un moment dans le temps. Selon lui, les évaluations « font partie de l'accumulation continue des connaissances sociales » (p. 318). Heureusement, distinguer des périodes temporelles n'impose pas de considérer l'utilisation comme événement, pas plus qu'elle n'exige de trancher le débat entre l'utilisation comme moment distinct ou comme processus. La dimension temporelle attire l'attention sur toute influence visible au cours d'une période donnée, qu'il s'agisse d'un événement qui ne se produise qu'au cours de cette période ou d'un processus qui la dépasse.

L'influence immédiate. L'influence immédiate désigne l'influence qui se produit ou qui est visible en même temps que le processus d'évaluation. Une influence immédiate peut survenir au cours du processus d'anticipation, de planification et de mise en œuvre de l'évaluation. Elle comprend les influences précoces qui sèment les graines d'effets à long terme ou qui peuvent avoir un impact cumulatif au fil du temps, ainsi que les effets à court terme qui peuvent ne pas avoir de ramifications à long terme. Bien que ce soient les promoteurs et promotrices de démarches participatives, collaboratives et d'autonomisation qui ont, chacun-e à sa manière, attiré notre attention sur cette question, l'influence immédiate n'est pas liée exclusivement à ces démarches.

Il suffit de voir, par exemple, l'influence de l'accréditation existante sur les préparatifs et le déroulement d'une visite d'évaluation sur site, avant même qu'une décision ait été prise quant à son renouvellement. Dans

un premier temps, l'influence immédiate peut être considérée comme exclusivement fondée sur le processus; toutefois, une réflexion approfondie sur la diversité des données qui constituent des résultats donne à penser que l'utilisation basée sur les résultats peut également se produire en même temps que le processus d'évaluation. L'étude d'évaluabilité, par exemple, indique clairement l'ordre du jour de la préparation des systèmes à évaluer (Wholey, 1994). Les ajustements apportés par un système en réponse aux données sur l'évaluabilité au cours du processus d'évaluation représentent une influence immédiate, intentionnelle et basée sur les résultats.

Deux clarifications s'imposent ici. Premièrement, l'influence immédiate ne se produit pas nécessairement sur un temps réduit. Étant donné qu'elle est liée au calendrier de l'évaluation, une mission d'évaluation qui s'étend sur une période de plusieurs mois ou même plusieurs années pourrait avoir une période prolongée au cours de laquelle l'influence immédiate pourrait être examinée. Deuxièmement, la caractérisation d'immédiate ne correspond pas à la durée de l'influence. On peut avoir une influence immédiate de courte durée ou une influence qui se prolongerait au-delà du cycle d'évaluation et qui demeurait visible au cours de périodes ultérieures.

L'influence en fin de cycle. La notion d'influence de fin de cycle met l'accent sur l'influence environnant la conclusion d'une étude d'évaluation sommative ou d'un cycle dans une évaluation plus formative. Elle comprend l'influence qui émane à la fois des produits de l'évaluation (par exemple, rapports, résumés et autres documents) et du processus de diffusion des résultats. Elle comprend également le processus qui met fin à un cycle d'évaluation particulier en l'absence d'un rapport écrit officiel et dans le contexte d'une utilisation plus poussée (Patton, 1994, 1997). L'influence de fin de cycle est parallèle à la notion d'effets de fin d'intervention dans l'évaluation des résultats, en attirant l'attention sur la conclusion d'une étude d'évaluation ou d'un cycle donné dans le cadre

d'un effort d'évaluation continu. Ces démarcations cycliques peuvent représenter des phases de développement de l'évaluation elle-même ou découler d'exigences du programme comme des cycles de financement. Comme l'a fait remarquer Patton (1997), la clôture d'un cycle peut inclure ou non un rapport d'évaluation, même si les marqueurs pour la fin d'un cycle peuvent être moins clairs en l'absence d'un tel produit.

Brett, Hill-Mead et Wu (dans le présent volume) fournissent des exemples particulièrement clairs de cycles dans un contexte plus large d'évaluation continue. Bien que l'on s'intéresse traditionnellement dans ce calendrier à l'examen de l'utilisation basée sur les résultats, il est aussi possible de s'intéresser à l'influence basée sur le processus. L'influence basée sur le processus au cours de cette période comprendrait les effets des interactions de réseautage entourant la finalisation, la clôture ou la liquidation d'une évaluation. Brett, Hill-Mead et Wu décrivent une influence de fin de cycle qui relie le processus et les résultats lorsqu'ils constatent que la réflexion structurée et basée sur les données du processus de synthèse trimestriel a appris au personnel à encadrer les membres du personnel lors de l'élaboration des objectifs pour l'année suivante. Une théorie intégrée de l'influence élargit également la focale pour observer des influences imprévues en fin de cycle qui peuvent découler soit du processus, soit des résultats.

L'influence à long terme. La notion d'influence à long terme permet de saisir des effets qui peuvent ne pas être ressentis pendant une période donnée ou qui évoluent au fil du temps en effets prolongés. L'inclusion explicite d'une utilisation future est utile pour rappeler aux évaluateurs et évaluatrices de ne pas s'arrêter à court terme dans leur examen de l'influence de leur travail. Bien que l'influence au cours du processus d'évaluation et de communication des résultats soit importante, l'impact le plus puissant du travail n'est peut-être pas encore apparu ou n'est pas visible dans ce délai, et il se situe plutôt dans un contexte futur. La conception de l'utilisation par Preskill et Torres comme apprentissage

transformateur (dans le présent volume) souligne l'importance d'une perspective à long terme, considérant l'apprentissage comme un processus continu de dialogue et de réflexion qui se produit progressivement au fil du temps. La première étape vers le suivi et l'étude empirique de l'impact futur est la reconnaissance de sa pertinence conceptuelle par rapport à une théorie intégrée de l'influence.

L'importance de comprendre l'influence à long terme a été soulignée sous l'angle théorique, éthique et pragmatique (Alkin, 1990; Shulha et Cousins, 1997). Bien que l'importance de l'influence à long terme soit bien reconnue (Huberman et Cox, 1990; Patton, 1986; Weiss, 1980), Shulha et Cousins (1997) ont trouvé qu'il était absent d'études empiriques sur l'utilisation.

Bien que les études de recherche aient rapporté – habituellement au moyen de méthodes relativement immédiates et rétrospectives – les conséquences instrumentales, cognitives, affectives et politiques de l'évaluation, elles ne suivent généralement pas ces dimensions à long terme. Par conséquent, elles ne permettent pas de dresser un tableau complet des changements personnels et professionnels chez les participant-e-s et des changements culturels dans les organisations (p. 204).

La notion d'influence à long terme reconnaît que l'influence peut être visible bien au-delà de la fin d'un cycle d'évaluation particulier. Elle invite les évaluateurs et évaluatrices à surveiller l'émergence d'impacts qui sont chronologiquement plus éloignés de l'évaluation et à identifier *a posteriori* des impacts antérieurs au fil du temps. L'influence à long terme peut être retardée, prolongée, ou les deux. En cas d'influence retardée, par exemple, les résultats de l'évaluation d'une étude sur la satisfaction des usagers et usagères peuvent au départ ne pas être utilisés en raison d'autres exigences du programme. Toutefois, quelques années plus tard, dans le cadre de la refonte du programme pour le prochain cycle de déploiement le personnel du programme peut reconnaître la pertinence des données déjà recueillies et les intégrer au bilan pour orienter la nouvelle programmation. Dans le cas d'une influence durable, les résultats

peuvent exercer une influence continue qui date du processus d'évaluation lui-même. Les données des *focus group* ont d'abord été utilisées pour fournir des retours du terrain aux opérateurs et opératrices du programme, puis incorporées dans le rapport annuel qui accompagne le cycle de financement du programme. La reddition de comptes est maintenue et le financement futur des programmes est assuré ou peut-être élargi. Dans cet exemple d'influence basée sur les résultats, les mêmes données qui exercent une influence à long terme ont déjà eu des effets immédiats et en fin de cycle. Dans un modèle combiné, une certaine influence à long terme se poursuit à partir de périodes antérieures et certaines émergent pour la première fois. Si, dans l'exemple précédent, les données des *focus group* étaient également utilisées dans les efforts de sensibilisation auprès des parties prenantes entrepris quelque temps après l'évaluation, cette influence retardée s'ajouterait à l'influence durable décrite précédemment.

Résumé. L'attention accordée à la temporalité de l'utilisation de l'évaluation n'est pas nouvelle, bien qu'historiquement les discussions sur le temps aient souvent été confondues avec d'autres dimensions plutôt qu'explicitement abordées. Les premiers travaux sur la dimension temporelle ont porté sur l'utilisation immédiate, et ce n'est que plus récemment que l'importance de l'utilisation à long terme a été soulignée. De même, les premières conceptualisations ont parlé de l'utilisation comme d'un événement, alors que les discussions plus récentes conçoivent l'évaluation comme un processus. La division de la dimension temporelle en termes immédiats, de fin de cycle et de long terme élargit la dichotomie commune entre court terme et long terme et s'inscrit en parallèle à la description chronologique conventionnelle des résultats du programme. Toutefois, la nature progressive de l'influence ne devrait pas être occultée par la démarcation de trois périodes. Il s'agit de tenir compte d'une gamme complète d'influences au fil du temps plutôt que de limiter la réflexion à une période étroite. C'est pourquoi la dimension

temporelle permet de prendre en compte à la fois le rythme du changement et les périodes chronologiques dans lesquelles il est mis en évidence. […]

Bibliographie

Alkin, Marvin C. 1990. *Debates on Evaluation*. Thousand Oaks: Sage Publications.

Alkin, Marvin C., Richard Daillak et Peter White. 1979. *Using Evaluations: Does It Make a Difference*? Thousand Oaks: Sage Publications.

Anderson, Christopher D., James A. Ciarlo et S. F. Brodie. 1981. « Measuring Evaluation-Induced Change in Mental Health Programs ». in *Utilizing Evaluation: Concepts and Measurement Techniques*, édité par J. A. Ciarlo. Thousand Oaks : Sage Publications, p. 97-123.

Anderson, Scarvia B., et Samuel Ball. 1978. *The Profession and Practice of Program Evaluation*. San Francisco: Jossey-Bass.

Argyris, Chris, Robert D. Putnam, et Diana McLain Smith. 1985. *Action Science*. San Francisco: Jossey-Bass.

Brisolara, Sharon. 1998. « The History of Participatory Evaluation and Current Debates in the Field ». in *Understanding and Practicing Participatory Evaluation*. Vol. New Directions for Evaluation, édité par E. Whitmore. San Francisco: Jossey-Bass, p. 25-41.

Caro, Francis G. 1980. « Leveage and Evaluation Effectiveness ». *Evaluation and Program Planning* 3(2) : 83-89.

Christie, Christina A., et Marvin C. Alkin. 1999. « Further Reflections on Evaluation Misutulization ». *Studies in Educationnal Evaluation* 25 : 1-10.

Ciarlo, James A. 1981. « Editor's Introduction ». in *Utilizing Evaluation: Concepts and Measurement Techniques*, édité par J. A. Ciarlo. Thousand Oaks: Sage Publications.

Cousins, J. Bradley, et Kenneth A. Leithwood. 1986. « Current Empirical Research on Evaluation Utilization ». *Review of Educational Research* 56(3) : 331-64.

Cousins, J. Bradley, et Elizabeth Whitmore. 1998. « Framing Participatory Evaluation ». *New Directions for Evaluation*, (80) : 5-23. doi : https://doi.org/10.1002/ev.1114

Cronbach, Lee J. 1982. Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass.

Cronbach, Lee J., Sueann R. Ambron, Sanford M. Dornbusch, Robert D. Hess, Robert C. Hornik, Denis Charles Phillips, Decker F. Walker, et Stephen S. Weiner. 1980. *Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements*. San Francisco: Jossey-Bass.

Fetterman, David M. 1994. « Empowerment evaluation ». *Evaluation Practice* 15(1) : 1-15. doi : https://doi.org/10.1177%2F109821409401500101

Greene, Jennifer C. 1988. « Communication of Resultas and Utilization in Participatory Program Evaluation ». *Evaluation and Program Planning* 11(4) : 341-51.

Greene, Jennifer C. 1988b. « Stakeholder Participation and Utilization in Program Evaluation ». *Evaluation Review* 12(2) : 91-116.

Henry, Gary T., et Debra J. Rog. 1998. « A Realist Theory and Analysis of Use ». in *Realist Evaluation: An Emerging Theory in Support of Practice.*, édité par G. T. Henry, G. Julnes, et M. M. Mark. San Francisco: Jossey-Bass.

Huberman, Michael, et Pat Cox. 1990. « Evaluation Use: Building Links Between Action and Reflection ». *Studies in Educational Evaluation* 16 : 157-79.

Johnson, R. Burke. 1998. « Toward a Theoretical Model of Evaluation Utilization ». *Evaluation and Program Planning* 21(1) : 93-110. doi: https://doi.org/10.1016/S0149-7189(97)00048-7

Kirkhart, Karen E. 1995. « Consequential Validity and an Integrated Theory of Use ».

Kirkhart, Karen E. 1999. « Multifaceted Dimensions of Use: Intended and Unintended Influences ».

Knorr, Karin D. 1977. « Policymakers' Use of Social Science Knowledge: Symbolic or Instrumental ? » in *Using Social Research in Public Policy Making*. Lexington Mass: Heath.

Leviton, Laura C., et Edward F. X. Hughes. 1981. « Research on the Utilization of Evaluations : A Review and Synthesis ». *Evaluation Review* 5(4) : 525-48. doi : https://doi.org/10.1177%2F0193841X8100500405

McClintock, Charles, et Laura A. Colosi. 1998. « Evaluation of Welfare Reform: A Framework for Addressing the Urgent and the Important ». *Evaluation Review* 22(5) : 668-94. doi : https://doi.org/10.1177/0193841×9802200505

Patton, Michael Q. 1986. *Utilization-Focused Evaluation*. 2e éd. Beverly Hills: Sage Publications.

Patton, Michael Q. 1994. « Developmental Evaluation ». *Evaluation Practice* 15(3) : 311-19.

Patton, Michael Q. 1997. *Utilization-focused evaluation: The new century text*. 3e éd. Thousand Oaks: Sage Publications.

Patton, Michael Q. 1998. « Discovering Process Use ». *Evaluation* 4(2) : 225-33.

Rein, Martin, et Sheldon H. White. 1975. « Can Policy Research Help Policy? » *Public Interest* 49 : 119-36.

Rich, Robert F. 1977. « Use of Social Sciences Information by Federal bureaucrats: Knowledge for Action Versus Knowledge for Understanding ». in *Using Social Research in Public Policy Making*, édité par C. H. Weiss. Lexington, Mass: Heath.

Rippey, Robert M. 1973. *Studies in Transactional Evaluation*. Berkeley: McCutchan.

Rodman, Hyman, et Ralph Kolodny. 1972. « Organizational Strains in the Researcher-Practitioner Relationship ». in *Evaluating Action Programs: Readings in Social Action and Education*, édité par C. H. Weiss. Needham Heights: Allyn & Bacon.

Scriven, Michael. 1991. *Evaluation Thesaurus*. 4e éd. Thousand Oaks: Sage Publications.

Shadish, William R., Thomas D. Cook et Laura C. Leviton. 1991. *Foundations of program evaluation: Theories of practice*. Newberry Park: Sage Publications.

Shulha, Lyn M., et J. Bradley Cousins. 1997. « Evaluation use: Theory, research, and practice since 1986 ». *Evaluation Practice* 18(3) : 195-208. doi: https://doi.org/10.1177%2F109821409701800302

Smith, M. F. 1988. « Evaluation Use Revisited ». J. A. McLaughlin, L. J. Weber, R. W. Covert et R. B. Ingle, dir. *Evaluation Utilization*. San Francisco: Jossey-Bass, p. 7-19.

Tornatzky, Louis G. 1979. « The Triple-Threat Evaluator ». *Evaluation and Program Planning* 2(2) : 111-15.

Weiss, Carol H. 1980. « Knowledge creep and decision accretion ». *Diffusion* 1(3) : 381-404.

Weiss, Carol H. 1981. « Measuring the Use of Evaluation ». in *Utilizing Evaluation: Concepts and Measurement Techniques*. Thousand Oaks: Sage Publications.

Weiss, Carol H., et Michael J. Bucuvalas. 1980. « Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Sciences Research ». *American Sociological Review* 45 : 302-13.

Whitmore, Elizabeth. 1991. « Evaluation and Empowerment: it's the Process That Counts ». *Empowerment and Family Support Networking Bulletin (Cornell University Empowerment Project)* 2(2) : 1-17.

Wholey, Joseph S. 1994. « Assessing the Feasibility and Likely Usefulness of Evaluation ». in *Handbook of Practical Program Evaluation*, édité par J. S. Wholey, H. P. Hatry, et K. E. Newcomer. San Francisco: Jossey-Bass.

5. Au-delà des usages : comprendre l'influence de l'évaluation sur les attitudes et les actions

GARY T. HENRY ET MELVIN M. MARK

[Traduit de : Henry, Gary T., Melvin M. 2003. « Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions ». *American Journal of Evaluation* 24(3) : 293-314 (Extraits). Traduction par Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Vers une théorie de l'influence de l'évaluation : cheminements et mécanismes multiples

Dans cette partie, nous présentons un cadre permettant de mieux comprendre l'influence de l'évaluation. Dans ce cadre, nous classons les processus de changement et les résultats que les évaluations peuvent influencer en fonction de trois niveaux : individuel, interpersonnel et collectif. Les niveaux indiquent le lieu du processus de changement. Par exemple, un processus de niveau individuel ne concerne qu'une seule personne, tandis qu'un processus de niveau interpersonnel se déroule dans le cadre d'un échange entre deux personnes ou plus, et qu'un processus de niveau collectif se déroule au sein d'une organisation ou d'une institution. À titre d'exemple, la lecture des conclusions d'une évaluation pourrait entraîner un changement : dans la pensée de Susan (individuel), dans les interactions entre Susan et Juan alors qu'elle tente de le persuader de quelque chose lié à l'évaluation (interpersonnel), ou dans une décision législative reflétée dans un vote majoritaire (collectif).

Après avoir décrit diverses formes d'influence de l'évaluation à ces trois niveaux, nous examinons la façon dont un processus ou un résultat peut en déclencher un autre, en passant parfois d'un niveau d'analyse à un autre (de l'individuel à l'interpersonnel au collectif). En d'autres termes, nous discutons des multiples cheminements alternatifs par lesquels les évaluations peuvent produire leurs effets.

Trois niveaux d'influence de l'évaluation

Ces trois niveaux correspondent à différents types de processus de changement qui ont été étudiés dans les sciences sociales et comportementales. Le premier niveau, individuel, désigne les cas où le processus ou les résultats de l'évaluation provoquent directement un changement dans les pensées ou les actions d'un ou de plusieurs individus, ou, plus précisément, lorsque le processus de changement ou les différences en question se produisent principalement au sein d'une personne. Par exemple, les évaluateurs et évaluatrices s'attendent souvent à ce que la prise de connaissance des résultats de l'évaluation change les croyances et les opinions des gens. Par ailleurs, la notion d'usage du processus suggère qu'un individu peut apprendre quelque chose ou changer ses croyances en fonction de sa participation à une évaluation (notez que la différence entre les conclusions de l'évaluation et le processus d'évaluation en tant que stimulateur du changement correspond à ce que Kirkhart (2000) désigne comme la source d'influence, et non pas au processus de changement lui-même. En prévision de notre discussion ultérieure sur les cheminements de l'évaluation, il est important de garder à l'esprit qu'un élément du cadre que nous présentons peut en stimuler un autre, que ce soit au sein des niveaux d'analyse, ou entre ces derniers. Par exemple, les évaluateurs et évaluatrices peuvent souvent supposer que le changement des attitudes individuelles entraînera des changements dans les comportements interpersonnels ou au niveau collectif.

Le deuxième niveau de notre cadre, l'interpersonnel, fait référence à un changement apporté dans les interactions entre les individus ou, plus précisément, à un processus ou un résultat qui se produit principalement dans le cadre des interactions entre les individus. Par exemple, les résultats d'une évaluation peuvent constituer une source faisant autorité, sur laquelle un intervenant ou une intervenante s'appuie pour tenter de changer les attitudes et les comportements d'autrui. Dans le contexte des programmes et de l'évaluation, le processus de persuasion est le plus souvent interpersonnel, un (ou plusieurs) individu s'efforçant d'influencer les autres. Le troisième niveau, le collectif, fait référence à l'influence directe ou indirecte de l'évaluation sur les décisions et les pratiques des organisations, qu'elles soient publiques ou privées. Plus précisément, le niveau collectif intervient lorsqu'un processus de changement ou un résultat opère principalement au sein d'un organisme social global. Par exemple, un changement de politique officielle pourrait être influencé par les conclusions d'une évaluation.

Pour orienter le lecteur ou la lectrice dans cette typologie, ces trois niveaux, ainsi que les formes spécifiques d'influence au sein de chaque niveau, sont résumés dans la figure 1. Pour chacun de ces niveaux, nous avons énuméré entre quatre et six types spécifiques de processus de changement ou de résultats qui pourraient vraisemblablement être déclenchés par une évaluation. Chacun de ces effets est issu de la littérature de recherche en sciences sociales et en sciences comportementales qui a examiné les processus de changement. En effet, la plupart des formes d'influence dont nous discutons ont des racines profondes dans les sciences sociales et se sont avérées fructueuses pour la recherche et les théories relatives aux les processus de changement. En spécifiant les différentes formes d'influence, nous avons largement emprunté à la recherche en psychologie, en science politique, en comportement organisationnel et dans d'autres domaines. L'un des facteurs qui a peut-être freiné la recherche et la théorisation des usages de l'évaluation est le fait que la littérature sur les usages a développé

une terminologie qui entrave plutôt qu'elle ne facilite l'intégration des connaissances provenant d'autres domaines qui examinent des processus de changement similaires.

Figure 1. Mécanismes par lesquels l'évaluation peut exercer une influence

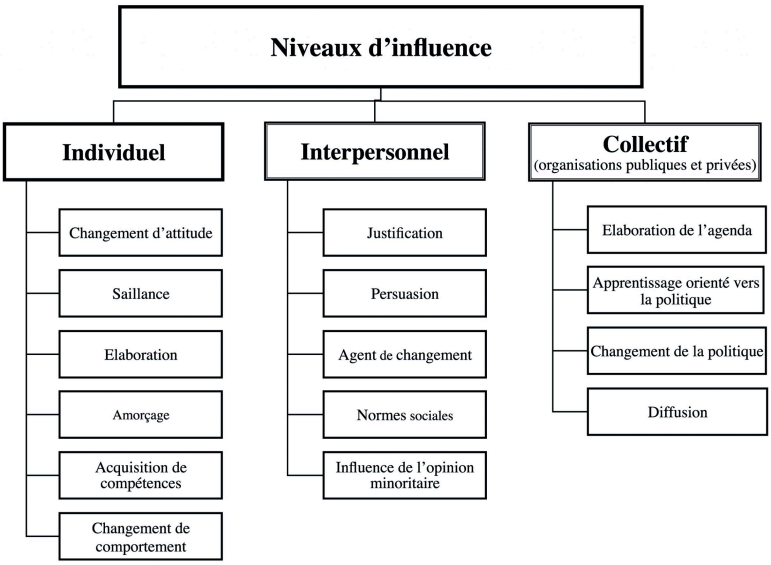


Tableau 1. Processus de changement pour l'influence de l'évaluation au niveau individuel

Mécanismes individuels	**Exemple**	**Références**
Changement d'attitude	Les résultats de l'évaluation influencent l'opinion des responsables politiques sur la faisabilité de la mise en œuvre du programme.	(Greenberg, Mandell, et Onstott, 2000)
Saillance	Les informations sur les effets du programme soulignent l'importance de la question ou du problème social.	(Henry et Gordon, 2001)
Élaboration	Le rapport d'évaluation incite les personnes à réfléchir davantage au programme et à leurs attentes quant à ses conséquences.	(Petty et Cacioppo, 1986)
Priorisation	La discussion des résultats des tests au début du rapport d'évaluation les rend plus importants dans le jugement du lecteur/de la lectrice sur l'efficacité du programme.	(Krosnick et Brannon, 1993)
Acquisition de compétences	De nouvelles compétences, telles que la collaboration ou les techniques d'enquête, sont acquises grâce à la participation à l'évaluation.	(King, 2002)
Changement de comportement	Les enseignant-e-s commencent à mélanger méthodes globale et phonétique après avoir appris que la combinaison des deux approches est plus efficace que l'une ou l'autre séparément.	(Patton, 1997)

Plusieurs points concernant la figure 1 méritent d'être soulignés.

Premièrement, nous ne nous attendons pas à ce que chaque étude sur l'influence tienne compte de tous les processus et résultats énumérés dans la figure 1, ni à ce que le plan de chaque praticien ou praticienne aborde tous les cheminements d'influence. La figure 1 propose plutôt un menu aux chercheurs et aux chercheuses, en utilisant des conceptions pré-ordonnées ou émergentes, et que les praticien-ne-s peuvent prendre en considération.

Deuxièmement, la liste des processus et des résultats de la figure 1, associée aux publications de sciences sociales pertinentes, devrait s'ajouter à ce qui est connu de la recherche et des savoirs courants sur les usages, avec des tests supplémentaires des facteurs liés à l'influence de l'évaluation.

Troisièmement, la prise en compte des processus de changement social que nous avons identifiés pourrait alimenter les discussions futures sur la formation en évaluation et les compétences des évaluateurs et évaluatrices.

Quatrièmement, il est vrai que nous avons fait des choix dans la sélection des éléments énumérés à la figure 1. Toutefois, nous pensons que le cadre résumé dans la figure pourrait stimuler les travaux susceptibles d'aboutir à de nouveaux développements et améliorations. Notre traitement des processus de changement est bref et est conçu pour appuyer notre discussion sur la façon dont ces résultats distincts peuvent être reliés en chaînes qui constituent des cheminements par lesquels les évaluations exercent une influence.

[…]

Les multiples cheminements d'influence de l'évaluation

Dans la pratique, des résultats spécifiques (tels que le changement d'attitude d'un individu et la persuasion interpersonnelle) seront souvent liés les uns aux autres comme les maillons d'une chaîne causale. Étant donné que l'influence d'une seule évaluation peut se manifester à travers de nombreuses chaînes de résultats, il existe de multiples cheminements d'influence possibles. Pour ne prendre qu'un exemple, un cheminement d'influence pourrait commencer par une personne qui apprend d'une évaluation multi site que le programme local ne produit pas les résultats escomptés et qui, par conséquent, change d'attitude à l'égard du programme local. Cette influence portant sur l'attitude au niveau individuel pourrait ensuite amener la personne à jouer le rôle d'agent de changement. À son tour, ce type spécifique de comportement interpersonnel pourrait, après un certain temps, conduire à l'adoption d'un modèle de programme plus efficace à partir d'un autre site de l'évaluation, c'est-à-dire à la diffusion. Mais il se pourrait qu'une fenêtre d'opportunité ait été ouverte dans l'agenda politique par d'autres évènements non liés, sans lesquels la politique n'aurait pas été prise en considération. L'évaluation est une influence potentielle à tous ces niveaux, mais l'objectif d'amélioration sociale peut être contrecarré par d'autres éléments de l'environnement qui neutralisent l'influence de l'évaluation ou, de façon plus bénigne, ne parviennent pas à créer un environnement propice à la progression de l'évaluation le long de la chaîne souhaitée.

Nous pourrions imaginer un cas où les résultats d'une évaluation arment un groupe dont les attitudes sont initialement minoritaires à l'égard d'un problème social existant et où le groupe fait connaître les résultats par l'intermédiaire des médias et lors de réunions avec des législateurs individuels, ce qui augmente par la suite l'importance perçue du problème social (c'est-à-dire sa mise à l'agenda). Notez que ce cheminement

particulier ne commence pas par un changement d'attitude. La chaîne de causalité commence plutôt par les conclusions de l'évaluation qui dynamisent le groupe d'opinion minoritaire et lui donnent un outil à utiliser dans ses efforts pour faire valoir ses préoccupations comme « problème public » (en faisant connaître les conclusions de l'évaluation pour tenter de susciter une préoccupation chez le public). Lorsque la question est inscrite comme « problème public », les membres du public peuvent commencer à réfléchir davantage à la question et commencer à élaborer leurs opinions, ce qui peut entraîner un changement d'attitude quant à l'opportunité d'une action publique sur la question. Enfin, une action collective se produit, sous la forme d'un changement de politique, qui, selon certaines recherches, peut en fait résulter de la mise à l'agenda d'un problème public et de l'élaboration d'une posture (Monroe, 1998; Page et Shapiro, 1992).

Pour emprunter un exemple légèrement en dehors de la sphère de l'évaluation, ce cheminement correspond approximativement à la façon dont des groupes tels que *Mothers Against Drunk Driving* [Les Mères contre l'alcool au volant] ont exercé une influence sur le changement de politique. À partir d'une position minoritaire (sur les activités d'application de la loi et les sanctions pénales applicables à la conduite en état d'ivresse), les membres du MADD se sont armées de statistiques (ainsi que d'exemples déchirants de décès). Leurs efforts en tant que groupe d'opinion minoritaire ont contribué à placer la question de la conduite en état d'ivresse au premier plan de l'opinion publique. Il est probable qu'en raison de la réflexion que cela a suscitée sur la question (voir la catégorie « élaboration »), un changement d'attitude s'est produit. Alors que le public considérait autrefois que la question de l'alcool au volant était un problème privé, le risque étant encouru par un individu, les opinions ont changé lorsque le comportement a été perçu comme mettant d'autres personnes en danger. Un changement de politique a suivi, sous la forme de lois plus sévères sur la conduite en état d'ivresse.

Remarquez que ce récit fournit beaucoup de détails et ne se contente pas de dire que « les statistiques sur les décès des accidents de la circulation liés à l'alcool ont été utilisées de façon instrumentale ». Ce genre d'analyse peut également permettre de comprendre pourquoi une évaluation particulière (ou, plus généralement, tout autre effort de changement) ne parvient pas à produire un type de changement. Imaginez, par exemple, que peu de temps après les efforts de MADD pour inscrire la conduite en état d'ivresse comme problème public, les États-Unis aient été impliqués dans une guerre. Cela aurait pu couper tout l'oxygène de la couverture médiatique et du discours public sur la conduite en état d'ivresse. Cela aurait-il été la faute de l'équipe d'évaluation, ou une lacune de l'évaluation? Nous pensons que non, bien que cela puisse sembler être le cas à partir d'un point de vue normatif sur les usages. Imaginez encore que la question ait fait partie de l'agenda médiatique pendant un certain temps, mais que les gens n'aient pas développé leurs réflexions à ce sujet ou n'aient pas changé d'attitude, comme cela semble s'être produit avec la question de la maltraitance des personnes âgées dans les programmes de soins de santé à domicile financés par le gouvernement fédéral (Cook *et al.*, 1983). Bref, la réflexion sur les cheminements d'influence de l'évaluation permet de comprendre et d'étudier quand et pourquoi une certaine utilisation finale se produit et quand et pourquoi elle ne se produit pas. La modélisation de la chaîne de causalité d'un processus de changement planifié, comme l'élaboration d'une théorie de programme pour un nouveau programme, peut aussi permettre aux évaluateurs et aux autres participants au processus d'évaluation d'évaluer la plausibilité du plan en premier lieu.

Dans l'exemple précédent, nous avons présenté une chaîne qui découle des résultats de l'évaluation plutôt que du processus d'évaluation (Patton, 1997), bien que les deux soient des points de départ viables pour l'influence de l'évaluation. L'influence liée au processus d'évaluation a initialement une portée plus limitée que l'influence initiale potentielle des conclusions de l'évaluation, car l'accès au processus est souvent plus limité que l'accès aux rapports d'évaluation. Autrement dit, une chaîne

déclenchée par le processus d'évaluation doit nécessairement commencer par les participant-e-s à l'évaluation, plutôt que par le public qui pourrait éventuellement lire ou entendre parler des conclusions de l'évaluation. Néanmoins, une chaîne déclenchée par le processus peut toujours être puissante, surtout lorsque des meneurs d'opinions ou des décideurs clés sont inclus. Par exemple, dans le cadre d'une évaluation participative, les enseignant-e-s pourraient apprendre à appliquer de nouvelles méthodes d'évaluation à la suite de leur décision d'évaluer les portfolios de travaux d'étudiant-e-s. Les enseignant-e-s pourraient décider d'appliquer leurs nouvelles compétences pour évaluer plus fréquemment le travail de leurs élèves et modifier en conséquence leurs pratiques pédagogiques. Si ces pratiques se traduisent par une plus grande réussite des élèves, ces pratiques d'évaluation pourraient faire partie des normes sociales de l'école et être adoptées par d'autres, notamment par les enseignant-e-s débutant-e-s qui arrivent dans l'école, avec une pression normative subtile encourageant ces mêmes pratiques. […]

Bibliographie

Cook, Fay L. *et al.* 1983. « Media and agenda setting: Effects on the public, interest group leaders, policy makers, and policy ». *Public Opinion Quarterly* 47 : 16-35.

Greenberg, David, Marvin Mandell et Matthew Onstott. 2000. « The dissemination and utilization of welfare-to-work experiments in state policymaking ». *Journal of Policy Analysis & Management* 19(3) : 367-82. doi : 10.1002/1520.

Henry, Gary T., et Craig S. Gordon. 2001. « Tracking issue attention: Specifying the dynamics of the public agenda ». *Public Opinion Quarterly* 65(2) : 157-77. doi : https://doi.org/10.1086/322198.

King, Jean A. 2002. « Building the evaluation capacity of a school district ». in *The art, craft, and science of evaluation capacity building*, édité par D. W. Compton, M. L. Baizerman et S. H. Stackdill. San Francisco: Jossey-Bass, p. 63-80.

Kirkhart, Karen. 2000. « Reconceptualizing evaluation use: An integrated theory of influence ». in *The expanding scope of evaluation use*, édité par V. J. Caracelli et H. S. Preskill. San Francisco: Jossey-Bass, p. 5-24.

Krosnick, Jon A., et Laura A. Brannon. 1993. « The impact of the gulf war on the ingredients of presidential evaluations: Multidimensional effects of political involvement ». *American Political Science Review* 87(4) : 963-75.

Monroe, Alan D. 1998. « Public opinion and public policy: 1980-1993 ». *Public Opinion Quarterly* 62(1) : 6-27.

Page, Benjamin, et Robert Y. Shapiro. 1992. *The rational public*. Chicago: University of Chicago Press.

Patton, Michael Q. 1997. *Utilization-focused evaluation: The new century text*. 3e éd. Thousand Oaks: Sage Publications.

Petty, Richard, et John Cacioppo. 1986. *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

6. Vers une évaluation post-normale?

THOMAS A. SCHWANDT

[Traduit de : Schwandt, Thomas A. 2019. « Post-normal evaluation? ». *Evaluation*, 25(3) : 317-329 (Extraits). Traduction de Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

[…] Je suis conscient du danger d'affirmer que l'on peut parler d'évaluation normale ou typique, étant donné la prolifération des démarches et pratiques d'évaluation dans le monde entier. Pourtant, les définitions officielles et les énoncés de mission des associations et sociétés d'évaluation, ainsi que les politiques et les cadres d'évaluation tels qu'ils ont été définis par les grandes organisations, notamment la Commission Européenne, la Banque mondiale, l'Agence des États-Unis pour le Développement International (USAID), le Ministère du Développement International du Royaume-Uni et le Groupe des Nations Unies pour l'Évaluation, font preuve d'une remarquable cohérence dans la définition de l'évaluation et du rôle qu'elle joue dans la société.

Du point de vue de la gouvernance, l'évaluation normale est une pratique qui s'intègre dans et qui promeut un mode de gouvernance axé sur l'intervention auprès des citoyens (*delivery mode of governance*) : les citoyennes et citoyens sont les destinataires, les bénéficiaires des interventions sociales, économiques et environnementales que leur adresse le gouvernement. L'évaluation vise à évaluer le succès de ces interventions en tant que solutions aux problèmes de la société. Stame et Furubo (2019) ont récemment décrit trois hypothèses centrales de l'évaluation normale, bien qu'ils n'aient pas utilisé cette appellation : (1) l'élaboration des politiques publiques se fait par le biais d'interventions ciblant des problèmes sociaux spécifiques (par opposition à la vision de

la gouvernance comme une lutte continue avec des problèmes complexes interdépendants) et l'évaluation est un outil permettant d'améliorer les interventions existantes, (2) l'environnement politique est supposé être relativement stable, et (3) l'évaluation des interventions passées est un atout clé pour acquérir des connaissances en vue d'interventions futures. Dans le cadre de son alliance avec les mécanismes de gouvernance, la pratique de l'évaluation s'est généralement engagée directement ou indirectement dans les idées du libéralisme ou du néolibéralisme, en établissant la valeur des solutions étatiques ou marchandes visant à résoudre des problèmes sociaux, économiques et environnementaux.

L'évaluation normale est liée aux notions de rationalité scientifique, de progrès social, d'efficacité et d'efficience des programmes sociaux, ainsi qu'à l'idéologie générale de la modernisation. Comme l'a souligné Peter Dahler-Larsen (2012), l'évaluation est véritablement moderne dans son ambition, dans sa croyance optimiste en la possibilité d'améliorer la société par la collecte de données et la prise de décisions rationnelles. Depuis ses débuts, l'évaluation normale s'est préoccupée des questions de rigueur, d'indépendance, de responsabilité, de cadres logiques (et leur raffinement connu sous le nom de théorie du changement), de résultats mesurables et d'aide à la minimisation des risques. Plus précisément, l'évaluation a généralement été considérée comme un moyen fiable d'assurer une certaine certitude dans l'appréciation de la valeur. Elle s'est enorgueillie de sa capacité à fournir des déterminations indépendantes, externes et expertes de la valeur en utilisant la logique unique de l'évaluation. Il s'agit d'établir une base de données probantes sur les meilleures pratiques qui aident à atteindre les buts et les cibles fixés tout en prenant en compte les effets intentionnels et non intentionnels des programmes et des politiques (bien que, entre parenthèses, je pourrais noter que les effets les plus importants de nos choix et de nos décisions sont précisément ceux qui sont considérés comme non intentionnels ou secondaires dans le langage d'Ulrich Beck).

Au risque de caricaturer (et en empruntant une métaphore utilisée récemment par l'un de mes collègues), l'évaluation normale a cherché à jouer le rôle de chien de garde et de chien-guide scientifique, en veillant à ce que la prise de décision démocratique prenne en compte des résultats valides et fiables quant à la valeur des politiques publiques et en produisant ces mêmes résultats. Elle a eu pour souci d'éviter de jouer le rôle de chien « à la botte », c'est-à-dire en résistant à la pression d'agir comme un partenaire soumis à la direction et aux commanditaires en ne leur apportant que des bonnes nouvelles sous un vernis scientifique.

L'évaluation normale est également, à quelques exceptions près, institutionnalisée et a acquis le statut d'industrie dirigée par de grands commissaires et prestataires (par exemple, le Groupe Indépendant d'Évaluation de la Banque Mondiale, l'Unité Indépendante d'Évaluation du Fonds vert pour le Climat, le Bureau Indépendant d'Évaluation du Programme des Nations Unis pour le Développement (PNUD), l'Unité Indépendante d'Évaluation de la Banque Asiatique de Développement, le Bureau Indépendant d'Évaluation du Fonds International pour le Développement Agricole). Cette institutionnalisation est également évidente dans le dispositif d'évaluation de la Commission Européenne, dans une variété d'efforts de renforcement des capacités d'évaluation parrainés par des fondations, des agences des Nations Unies et d'autres organisations; ainsi que dans le cadre des efforts visant à formuler des politiques d'évaluation gouvernementales, comme le préconisent, par exemple, l'*American Evaluation Association's Evaluation Policy Initiative* [Initiative sur les politiques d'évaluation de l'Association américaine d'évaluation] et le *Global Parliamentarian Forum for Evaluation* [Forum parlementaire mondial pour l'évaluation].

Une évaluation post-normale?

Les caractéristiques du *zeitgeist* (esprit du temps) identifiées précédemment remettent en question cette conception normale de l'évaluation. En d'autres termes, ma question est la suivante : « Les schémas directeurs de la modernité pour la théorie et la pratique de l'évaluation sont-ils épuisés? ». Assistons-nous peut-être à l'émergence d'une « évaluation post-normale? ». J'ai emprunté l'idée d'évaluation post-normale à la littérature sur la science post-normale : le concept de post-normal a été introduit pour la première fois par Jerôme Ravetz, le célèbre philosophe des sciences britannique, et le mathématicien argentin Silvio Funtowicz (Funtowicz et Ravetz, 1993). Ils ont fait valoir que l'ancienne image de la science, où les données empiriques aboutissaient à des conclusions incontestables et où le raisonnement scientifique aboutissait à corriger les politiques publiques, n'était plus plausible. En d'autres termes, la science ne fonctionnait plus de manière « normale ». Ils notent « chaque fois qu'un enjeu de politique publique implique la science, nous découvrons que les faits sont incertains, que la complexité est la norme, que les valeurs sont contestées, que les enjeux sont élevés, que les décisions sont urgentes et qu'il existe un réel danger que les risques causés par l'homme échappent à tout contrôle » (p. 737).

Ils ont décrit ces développements émergents en termes de « science post-normale », désormais un domaine d'étude établi. Les philosophes et spécialistes des sciences sociales ont étendu la science post-normale à l'ensemble de la société en utilisant l'expression « époque post-normale » et en soulignant la complexité, le chaos et les contradictions comme caractéristiques de cette époque (voir Sardar, 2010).

Avant de spéculer sur les grandes lignes de la théorie et de la pratique de l'évaluation post-normale, je propose quelques mises en garde : premièrement, je ne parle pas d'une révolution dans la théorie et la pratique de l'évaluation. Plus simplement, je suggère qu'un remaniement important de la théorie et de la pratique de l'évaluation pourrait être

en cours, et que d'autres pourraient encore survenir. Comme l'a souvent dit Sherlock Holmes (en citant Shakespeare), « le jeu est en marche ». Le terrain montre déjà quelques signes de pensée post-normale, pour ainsi dire. L'intérêt croissant pour l'évaluation du développement en est un exemple, tout comme l'influence des idées issues de la pensée systémique et de la science de la complexité sur la pratique de l'évaluation, comme en témoignent le nombre croissant de publications consacrées à ce sujet, en particulier dans cette revue (par exemple Gerrits et Verweij, 2015; Larson, 2018). Le document récemment publié « *Principles for Effective Use of Systems Thinking in Evaluation Practice* »[1] [Principes pour une utilisation efficace de la réflexion en termes de systèmes dans les pratiques d'évaluation] du *Systems in Evaluation Topical Interest Group* [Groupe d'intérêt thématique sur les systèmes en évaluation] de *l'American Evaluation Association* en est un autre exemple. D'autres signes de la pensée post-normale incluent l'initiative d'évaluation Sud-Sud lancée en 2017 qui vise à remédier aux asymétries de longue date dans les pratiques mondiales d'évaluation, ainsi que l'initiative du Comité d'aide au développement (CAD) de l'Organisation de coopération et de développement économiques (OCDE), qui parraine un examen des critères du CAD en vigueur depuis longtemps, bien que l'on puisse se demander si cette dernière évolution mène à une pensée post-normale (voir, par exemple, le blog de Zenda Ofir « *Evaluation for Development* »[2] [l'évaluation pour le développement].

Deuxièmement, il existe de multiples communautés d'évaluation, par exemple l'évaluation en matière de développement, l'évaluation environnementale, l'évaluation axée sur la gouvernance, l'évaluation engagée vis-à-vis des savoirs autochtones, etc. Chacune de ces communautés pourrait bien trouver une importance ou un intérêt différent dans les aspects de l'évaluation post-normale, comme nous le verrons plus loin.

1. https ://www.systemsinevaluation.com/principles-for-systems-thinking/
2. zendaofir.com

Enfin, il est peu probable que l'évaluation post-normale marque une rupture radicale avec l'évaluation normale. En empruntant et en étendant l'idée de Baumann (2000) selon laquelle la modernité a une nature ambivalente et double, je crois que nous pourrions dire la même chose de la pratique de l'évaluation. D'une part, elle se caractérise en grande partie par un besoin d'ordre, pour définir, apprécier et rationaliser le monde pour qu'il soit contrôlable, prévisible et compréhensible. Cette tendance à l'ordre et à la rationalisation est la force caractéristique de la modernisation. D'autre part, l'évaluation normale est, depuis un certain temps, également caractérisée par le changement, par de nouvelles façons de penser qui remettent en question cette tendance de rationalisation et qui critiquent fortement et cherchent à renverser les pratiques et les formes traditionnelles d'évaluation, comme cela est peut-être plus évident dans certaines formes d'évaluations transformatives et sensibles aux différences culturelles.

Indices d'une évaluation post-normale

La première indication est vraiment une question sur l'idée même d'innovation en évaluation, à savoir : « la stratégie a-t-elle remplacé la substance dans le discours en évaluation? ». C'est-à-dire, les principales préoccupations du domaine de l'évaluation portent-elles vraiment sur la manière dont les évaluateurs font leur travail, ou bien plutôt sur ce en quoi ce travail consiste? La justification centrale de l'évaluation – l'établissement de la valeur – ne devrait-elle pas signifier quelque chose qui ne soit pas seulement quantitatif et instrumental, mais qui soit qualitatif et substantiel? Définir la valeur ne devrait-il pas contribuer de manière significative et directe aux débats sur l'avenir de la société, plutôt que de se résumer à une question de contrôle et de jugement de la performance des interventions? L'innovation en matière d'évaluation se limite-t-elle vraiment à la façon dont le travail est effectué, à la méthode? Un récent webinaire sur les innovations en matière d'évaluation parrainé

par *Better Evaluation* [Meilleure Évaluation], le Fonds des Nations Unies pour l'enfance (UNICEF) et EVALSDG en est un exemple. La promotion du webinaire affirmait :

> Lorsque les outils, les processus, les méthodes et les systèmes d'évaluation existants ne suffisent pas, vous devez tirer parti des innovations en matière d'évaluation, et les innovations discutées étaient la théorie négative du programme [negative program theory], les modèles logiques à trois rangées, la répétition des données [data rehearsal], le big data, la collecte, la désagrégation et l'établissement de rapports par sous-groupes, et les rubriques.

Je ne dis pas que l'innovation technique n'est pas nécessaire, mais seulement qu'elle est loin d'être suffisante pour relever les défis actuels. Les deux indications suivantes concernent la façon dont nous comprenons la notion de résilience.

Penser une gouvernance fondée sur la résilience

La résilience est généralement définie comme la capacité des individus et des sociétés à se rétablir, à s'adapter et à prospérer face aux difficultés. Mais David Chandler (2014b) dans son livre *Resilience: The Governance of Complexity* [Résilience : La Gouvernance de la Complexité], et dans des articles connexes (Chandler, 2014a) conçoit la résilience comme un mode de pensée et d'action fondé sur une ontologie de la complexité émergente. L'idée que la vie, le fait d'exister, est un phénomène complexe, relationnel, intégré et contextuel.

Cette perspective sur le monde est en contradiction avec une ontologie moderniste qui cherche à le comprendre au regard de ses entités, de leurs propriétés et de leurs relations. Sans complexité émergente, pour Chandler (2014a), il n'y a pas besoin de résilience. Selon lui,

> le raisonnement en termes de résilience ne devrait pas être compris de manière étroite, comme un simple renforcement des capacités des individus et des sociétés… mais, plus largement, comme une conception de la gouvernance qui s'éloigne de la compréhension moderniste d'un gouvernement agissant de manière instrumentale dans un monde où les politiques publiques pourraient se réduire à de simples relations de cause à effet » (p.58).

Selon cette conception de la résilience, la gouvernance passe d'un mode d'intervention (*delivery mode*) – avec des politiques instrumentales axées sur l'offre et les objectifs – à un mode relationnel – une gouvernance fondée sur la compréhension des processus et des capacités qui existent déjà et sur la manière dont ils peuvent être intégrés dans la conception des politiques publiques. Cela marque un passage de l'évaluation normale, fondée sur l'idée d'observations et de mesures indépendantes, à l'évaluation en tant qu'intervention systémique : une action visant intentionnellement à produire du changement, en relation avec une réflexion sur les limites (Midgley, 2000). Ces limites déterminent quelles observations empiriques et quelles considérations de valeur peuvent être considérées comme pertinentes et lesquelles sont laissées de côté ou considérées comme moins importantes. L'évaluation post-normale génère des connaissances concrètes et pratiques et est donc plus proche d'une orientation de recherche-action que d'une évaluation indépendante générée par des experts (Chandler, 2014b).

Le retour de la politique au peuple

L'évaluation normale est associée à l'idée d'un discours démocratique se déroulant dans la sphère politique formelle de la prise de décision. Les savoirs issus de l'évaluation alimentent et informent le processus de prise de décision et la formation de la volonté démocratique collective

(Chandler 2014b; voir aussi Dzur 2008). C'est l'hypothèse qui sous-tend l'objectif souvent cité de l'évaluation au service du progrès social, ainsi que les efforts visant à renforcer les capacités d'évaluation nationales, comme le font, par exemple, EvalPartners et le PNUD.

En revanche, dans une conception relationnelle de la gouvernance (par opposition à un mode d'intervention), la politique revient au peuple, à la sphère des pratiques, des interactions et des représentations quotidiennes (Chandler, 2014b). On bascule vers une conception de la démocratie comme un mode de vie – comme le disait John Dewey – plutôt que comme ensemble des caractéristiques de la politique formelle. C'est le domaine du travail public. Il est « public » au sens où il est visible, ouvert à l'inspection, et son importance est largement reconnue. Il s'agit d'un travail civique coopératif d'« un public : un mélange de personnes dont les intérêts, les origines et les ressources peuvent être très différents » (Dzur, 2008). La notion de travail public va au-delà de l'idée de délibération publique. Il suscite l'espoir que les citoyen-ne-s agissent en tant que co-créateurs et co-créatrices d'un monde public, en produisant ensemble une vie publique (Dzur, 2008). À mesure que la politique revient au peuple, la position professionnelle indépendante caractéristique de l'évaluation normale cède la place à un rôle d'évaluateur/-trice en tant que facilitateur/-trice du débat public de manière à partager le pouvoir et la responsabilité avec les citoyens (Schwandt 2017, 2018). Le politiste Albert Dzur (2008 : 130) définit ce rôle en termes de « professionnalisme démocratique » : [il s'agit de] « Partager des tâches auparavant professionnalisées et encourager la participation des profanes de manière à améliorer et à permettre un engagement public plus large et une délibération sur les grandes questions sociales à l'intérieur et à l'extérieur des domaines professionnels ».

Retrouver un raisonnement pratique

Le retour de la politique au peuple a pour corollaire le retour de l'importance du raisonnement pratique comme antidote à l'acceptation inébranlable du raisonnement scientifique dans l'élaboration des politiques publiques. Une des caractéristiques de notre époque est l'acceptation d'une rationalité qui s'appuie sur les sciences naturelles et l'économie, à l'exclusion de l'histoire, de la culture et de la politique, en bref, une rationalité scientifique « déconnectée de l'expérience, de l'empathie et de tout fondement moral » (Sarewitzd, 2011). Lorsque le raisonnement pratique est compris comme étant situé au sein des communautés politiques, lorsque le raisonnement n'est pas considéré en termes de rationalité technocratique et bureaucratique, mais comme étant inséparable de la façon dont nous considérons les gens et ce qu'ils font dans la vie quotidienne et dans la vie publique (Edmonson et Hüsler, 2012), nous comprenons alors comment le raisonnement scientifique doit intégrer la rationalité telle qu'elle est comprise dans les contextes sociaux, moraux et historiques. Le raisonnement pratique est précisément ce qui est nécessaire dans les situations où des décisions ou des actions sont requises, mais les connaissances sont en constante évolution et insuffisantes, et où les attentes des experts et de la vie quotidienne ne sont pas en mesure de fournir une orientation claire. Les problèmes que l'évaluation tente d'aborder à sa manière unique n'ont pas de solutions toutes faites et exigent des réponses combinant des capacités cognitives, émotionnelles, sociales, politiques et morales (Edmonson, 2012). L'évaluation post-normale alimente le raisonnement pratique plutôt que la rationalité scientifique. L'évaluation post-normale ne vise pas à devenir une science ou une technologie encore plus forte, mais un exercice de phronesis – c'est-à-dire un moyen moral et pratique de délibérer et d'agir en rapport avec des questions de fond dans la vie quotidienne.

Coproduction

Un cinquième indice d'une évolution vers l'évaluation post-normale, étroitement lié aux deux précédentes, est la prise de conscience de l'importance du phénomène connu sous le nom de coproduction, introduit initialement par la regrettée Eleanor Ostrom (1996). Certes, ce terme a de multiples significations dans l'élaboration des politiques publiques, la gouvernance et la recherche dans différents domaines. Dans le domaine des soins de santé, par exemple, la coproduction est décrite comme un moyen de travailler ensemble à l'amélioration de la santé et de créer des services de santé pilotés par les utilisateurs et centrés sur les personnes. Au Royaume-Uni, la coproduction est devenue un terme courant dans le discours du gouvernement et des politiques publiques. Je m'intéresse à la coproduction en tant que moyen de redéfinir la relation entre les professionnels du service public et les citoyens. La mutualité et la réciprocité remplacent la relation de dépendance du bénéficiaire ou de l'usager-e vis-à-vis de l'expert-e. Les citoyen-ne-s ne sont pas de simples bénéficiaires de services fournis par des expert-e-s, mais aussi des collaborateurs et collaboratrices qui apportent leurs connaissances, leur expérience, leurs compétences et leurs capacités à la création d'innovations sociales. Elle s'oppose à une méthode de prestation de services dans laquelle les citoyen-ne-s consomment des services publics conçus et fournis par les pouvoirs publics. La coproduction est également un espace d'exploration et un processus génératif qui conduit à des formes différentes, et parfois inattendues, de connaissances, de valeurs et de relations sociales. Ces processus dynamiques prennent la forme d'interactions entre les individus et les services, et impliquent des interactions entre les différentes logiques de participation et d'agendas politiques, entre différents modes de production de connaissances et entre différents types de valeurs (par exemple, l'économie, l'équité, la justice sociale; Filipe, Renedo, et Marston, 2017).

Responsabilité éthique

Le dernier marqueur de l'évaluation post-normale est la découverte du rôle central de l'éthique dans l'action et la production de connaissances. Il ne s'agit pas ici de l'interrelation bien connue entre la politique et l'éthique, que l'on retrouve dans le sens de l'obligation démocratique de la profession « de mener des évaluations défendables, justes et équitables dans des contextes politiques où les valeurs et les intérêts des différentes parties prenantes sont souvent en conflit » (Simons, 2006 : 255), ni de l'accent habituel sur les vertus éthiques dont le professionnel doit faire preuve (l'honnêteté, l'intégrité, etc.). Je ne m'intéresse pas non plus ici à la question familière de savoir si la production de connaissances scientifiques est, de quelque manière que ce soit, soumise à des jugements de valeur morale, sociale ou politique. Cela renvoie essentiellement à la question de longue date de la neutralité des valeurs et du problème qui se pose lorsque les pratiques se revendiquant de l'évaluation et de la science deviennent des formes d'activisme politique ou de défense des intérêts. Ce qui m'intéresse, ce sont les considérations éthiques inhérentes à la production de connaissances en évaluation.

Pour reprendre une idée défendue par feu Paul Cillers (2005) et par d'autres, accepter une ontologie de la complexité émergente transforme la façon dont nous pensons la connaissance. La considération éthique liée à la production de connaissances est la suivante : le fait de ne pas reconnaître l'incertitude et la complexité n'est pas simplement une erreur technique, c'est aussi une erreur éthique. C'est un échec éthique lorsque nous n'assumons pas la responsabilité de notre savoir. L'éthique n'est pas quelque chose qui vient s'ajouter à notre compréhension du monde. L'éthique fait déjà partie intégrante de ce que nous faisons.

Les partisans et partisanes de l'analyse en termes de systèmes l'ont dit à maintes reprises, mais il convient de le répéter : on ne peut pas étudier une situation dans son intégralité. Il faut choisir ce qui est inclus et ce qui est exclu, ce qui est important et ce qui ne l'est pas. Nous « cadrons »

(*frame*) une situation d'une certaine façon afin de la comprendre. Un cadre est une perspective qui permet d'aborder une situation d'une manière particulière. Le cadrage implique de fixer des limites et donc d'imposer des limites à notre compréhension. Et lorsque nous avons affaire aux limites de notre compréhension, nous avons affaire à l'éthique (Cillers, 2005 : 261). Puisque des limites doivent être fixées, l'évaluateur ou l'évaluatrice, en tant que professionnel-le, assume la responsabilité des conséquences de l'établissement de limites d'une manière particulière. Comme l'a expliqué Werner Ulrich, les évaluatrices et évaluateurs professionnel-le-s et les chercheurs et chercheuses en sciences sociales ne peuvent justifier leurs affirmations sur le cadrage et la fixation de limites sur la base de leur expertise théorique et méthodologique. En effet, si tel était le cas, nous céderions l'autorité à ces professionnel-le-s, car ils et elles doivent certainement savoir mieux que les gens ordinaires ce qui est le mieux pour tout le monde. Pour ce qui est d'obtenir et d'offrir des jugements sur les limites, les évaluateurs et évaluatrices n'ont aucun avantage de principe sur les citoyen-ne-s ordinaires. Ce qu'ils et elles ont, c'est une obligation éthique, en tant que professionnel-le-s, de poursuivre une « approche autoréflexive, autocorrective et auto-limitative de l'enquête… qui remet en question toutes les sources imaginables de tromperie », par exemple, dans ses présuppositions, ses méthodes et ses procédures, ses résultats et la traduction de ces résultats en affirmations évaluatives (Ulrich, 2017). […]

Bibliographie

Baumann, Zygmunt. 2000. *Liquid Modernity*. Cambridge: Policy Press.

Chandler, David. 2014a. « Beyond neoliberalism: Resilience, the new art of governing complexity ». *Resilience: International Policies, Practices and Discourses* 2:47-63. doi : https://doi.org/10.1080/21693293.2013.878544.

Chandler, David. 2014b. *Resilience: The Governance of Complexity*. London: Routledge.

Cillers, Paul. 2005. « Complexity, deconstruction and relativism ». *Theory, Culture & Society* 22(5):255-67. doi : https://doi.org/10.1177%2F0263276405058052.

Dahler-Larsen, Peter. 2012. *The Evaluation Society*. Stanford: Stanford University Press.

Dzur, Albert W. 2008. *Democratic Professionalism*. University Park: University of Pennsylvania Press.

Edmonson, Ricca. 2012. « Practical reasoning in place tracing "wise" inferences in everyday life ». in *Politics of Practical Reasoning*, édité par R. Edmonson et K. Hüsler. Lanham: Lexington Books, p. 111-30.

Edmonson, Ricca, et Karlheinz Hüsler, éd. 2012. *Politics of practical reasoning*. Lanham: Lexington Books.

Filipe, Angela, Alicia Renedo et Cicely Marston. 2017. « The co-production of what? Knowledge, values, and social relations in health care ». *PLoS Biology* 15(5). doi : https://doi.org/10.1371/journal.pbio.2001403.

Funtowicz, Silvio O., et Jerome R. Ravetz. 1993. « Science for the post-normal age ». *Futures* 31 : 735-55.

Gerrits, Lasse, et Stephan Verweij. 2015. « Taking stock of complexity in evaluation: A discussion of three recent publications ». *Evaluation* 21(4) : 4821-4491. doi : https://doi.org/10.1177%2F1356389015605204.

Larson, Ann. 2018. « Evaluation amidst complexity: Eight evaluation questions to explain how complex adaptive systems affect program impact ». *Evaluation* 24(3) : 353-62. doi : https://doi.org/10.1177/1356389018781357.

Midgley, Gerald. 2000. *Systemic Intervention*. New York: Springer.

Ostrom, Elinor. 1996. « Crossing the great divide: Coproduction, synergy, and development ». *World Development* 24(6) : 1073-87. doi : https://doi.org/10.1016/0305-750X(96)00023-X.

Sardar, Ziauddin. 2010. « Welcome to postnormal times ». *Futures* 42 : 435-44.

Sarewitzd, Daniel. 2011. « Liberalism's modest proposal or, the tyranny of scientific rationality ». *Breakthrough Journal*.

Schwandt, Thomas A. 2017. « Professionalization, Ethics, and Fidelity to an Evaluation Ethos ». *American Journal of Evaluation* 20(10) : 1-8. doi : https://doi.org/10.1177/1098214017728578.

Schwandt, Thomas A. 2018. « Acting together in determining value: A professional ethical responsibility of evaluators ». *Evaluation* 24(3) : 306-17. doi : https://doi.org/10.1177%2F1356389018781362.

Simons, Helen. 2006. « Ethics in evaluation ». in *The SAGE handbook of evaluation*, édité par I. F. Shaw, J. C. Greene et M. M. Mark. Thousand Oaks: Sage Publications, p. 243-65.

Stame, Nicoletta, et Jan Eric Furubo. 2019. « Preface ». in *The Evaluation Enterprise*, édité par J. E. Furubo et N. Stame. New York: Routledge, p. 14-25.

Ulrich, Werner. 2017. « If systems thinking is the answer, what is the question? Discussions on research competence (Expanded and updated version of Working Paper No.22, Lincoln School of Management, University of Lincoln, Lincoln, UK, 1998) ».

7. L'évaluation est-elle obsolète dans un monde de post-vérité?

ROBERT PICCIOTTO

[Traduit de : Picciotto, Robert. 2018 « Is evaluation obsolete in a post-truth world? ». *Evaluation and program planning*, 73 : 88-96 (Extraits). Traduction de Carine Gazier et Agathe Devaux-Spatarakis; traduction et reproduction du texte avec l'autorisation de Elsevier. La traduction de ce texte en français a été révisée par son auteur.]

L'évaluation entre en scène

Il est certain que les démagogues exploiteront toujours la colère et la frustration populaire dans leurs efforts pour tromper les électrices et électeurs. Mais on peut leur résister. Les biais cognitifs peuvent être surmontés. La déconstruction peut être déconstruite. L'analyse objective des politiques publiques peut orienter les prises de décisions. Plutôt que de déplorer l'avènement de l'ère de post-vérité, il est temps de s'attaquer à ses causes profondes : malheureusement, les décideurs et décideuses sont bien conscient-e-s du fait que, dans une démocratie libérale, les électrices et électeurs peuvent être induit-e-s à juger les politicien-ne-s en fonction des objectifs qu'ils et elles poursuivent et des résultats qu'ils et elles promettent, plutôt qu'en fonction de la véracité de leurs déclarations.

Les évaluateurs et évaluatrices peuvent rectifier cette situation en exposant les contradictions et les effets néfastes des décisions motivées par des intérêts particuliers. L'évaluation est tout à fait en mesure de responsabiliser l'autorité et de renforcer la transparence des processus de décision. Elle peut confronter les mensonges et les absurdités

endémiques et dévoiler la fourberie des communications astucieuses utilisées par les politicien-ne-s pour manipuler et contrôler les électeurs et électrices en faisant appel à leurs émotions et à leurs préjugés. De cette façon, en temps voulu, le bon sens et la réalité pourraient bien reprendre le dessus (Davis, 2017).

L'évaluation peut contribuer à accélérer ce processus. Elle a pour mission de guider les décideurs et décideuses vers des solutions basées sur des faits et des analyses objectives de l'expérience. Les évaluations indépendantes, sans retenue, disent la vérité au pouvoir. Certes, les évaluateurs et évaluatrices ne peuvent pas à eux et elles seul-le-s s'attaquer à tous les problèmes qui ont donné naissance au phénomène de la post-vérité, mais ils et elles ont le devoir de lui résister. Ils et elles peuvent faire équipe avec les chercheurs et chercheuses en sciences sociales pour exposer et contester le discours hégémonique des spécialistes de la post-vérité. À cette fin, ils et elles peuvent contribuer à la promotion de politiques publiques plus efficaces sur la base de données factuelles.

Ainsi, plutôt que d'être obsolète, l'évaluation est plus nécessaire que jamais dans un environnement opérationnel particulièrement exigeant. Toutefois, pour s'acquitter de sa mission et accroître sa portée, elle doit améliorer sa pratique, raffiner ses outils et faire face aux défis sociaux les plus urgents de notre époque. À cette fin, les objectifs, les commanditaires et les objets de l'évaluation devraient être réexaminés afin de relever le défi de la post-vérité et la communauté de l'évaluation devrait réviser ses orientations stratégiques. Enfin, elle devrait faire le nécessaire pour devenir une profession.

Nouveaux objectifs

La plupart des évaluations font le bilan des interventions par rapport aux objectifs visés. Elles négligent souvent les objectifs de développement durable (ODD) universellement approuvés qui pourtant nécessitent une évaluation systématique. Il s'agit notamment de l'éradication de la pauvreté, de l'autonomisation des femmes, de la qualité de l'éducation, de la sécurité alimentaire, de l'accès universel à l'eau potable et à l'assainissement, de l'énergie durable, de la gestion efficace des ressources naturelles, de la bonne gouvernance, de la paix sociale, de la réduction des inégalités et de la création d'un environnement mondial propice à un développement équitable et inclusif. Compte tenu de la multiplicité des menaces qui pèsent sur le bien-être des populations, des injustices sociales, et de la prévalence des conflits transfrontaliers, des catastrophes naturelles, du changement climatique, etc., il faudra adopter une large gamme d'approches évaluatives au niveau des politiques globales.

Nouveaux commanditaires

Les processus d'évaluation devraient être conçus de manière à relier et à informer une grande diversité d'acteurs et d'actrices économiques et sociaux. La discipline de l'évaluation, trop attachée à l'évaluation des interventions publiques traditionnelles, n'a pas suivi le rythme des transformations profondes du monde social et doit désormais s'adapter au rythme dynamique des nouveaux protagonistes de tous les secteurs sociaux.

Une révolution néo-libérale sans précédent a déferlé sur l'économie internationale depuis le début du siècle. Le nombre d'entreprises multinationales est passé de 7 300 à la fin des années 1960 à 65 000

au début du siècle et à au moins 100 000 à la fin de l'année 2010. Avec l'expansion du nombre de milliardaires, la mobilisation des ressources financières privées à des fins sociales s'est intensifiée.

Les fondations caritatives sont devenues actives dans l'aide internationale et le secteur public a cédé une bonne partie de son pouvoir à la « révolution associative ». Les organisations non gouvernementales sont maintenant profondément impliquées dans la prestation de services sociaux, tandis que des activistes interconnectés à l'échelle mondiale sont de plus en plus en mesure de façonner les règles du jeu de la vie économique et sociale. La conséquence est un déficit d'évaluation lié au rôle accru du secteur privé et de la société civile dans la sphère sociale.

Nouveaux objets

Des cocktails de méthodes d'évaluation sur mesure devront être mobilisés pour décortiquer les effets sociaux complexes des instruments d'action publique contemporains :

- Le financement de l'impact social mobilise des capitaux et des compétences privées dans la poursuite de l'inclusion sociale et de l'environnement. Comment vérifier que faire le bien et le faire bien sont compatibles?
- Les partenariats public-privé ont mis en lumière la négligence de l'intérêt public dans la répartition contractuelle des risques et des bénéfices entre les entreprises privées et les citoyen-ne-s.
- Investissements à impact social : les indicateurs souvent simplistes incorporés dans les systèmes de gestion des entreprises devraient être remplacés par des objectifs triples qui intègrent les coûts et les bénéfices économiques, sociaux et environnementaux.
- L'organisation de compétitions pour susciter des initiatives sociales innovatrices augmente rapidement, mais leur évaluation laisse

beaucoup à désirer.

- L'éventail complexe de produits d'assurance, des garanties de marché, d'initiatives financées par la « diaspora » et des transferts de fonds des migrant-e-s nécessitent une évaluation systématique.
- La responsabilité sociale des entreprises : l'augmentation des investissements directs étrangers génère une demande croissante d'évaluation objective.

Nouvelles orientations de la politique d'évaluation

Trois réorientations politiques caractérisent le paradigme d'une évaluation conçue pour réagir plus efficacement contre un monde de « post-vérité ». Premièrement, une internationalisation de l'évaluation. Deuxièmement, la diversification des utilisateurs et utilisatrices et des produits. Troisièmement, des méthodes et des processus d'évaluation plus souples, plus réactifs, et plus numériques.

Internationalisation

Au fur et à mesure que le centre de gravité de l'économie mondiale continue d'évoluer vers le Sud et l'Est, l'évaluation devra opérer au-delà des frontières. Elle devra devenir plus « internationale dans le sens où elle sera à la fois plus indigène, plus globale et plus transnationale » (Chelimsky et Shadish, 1997). Le processus est en cours. Au tournant du siècle, il n'existait que 20 associations d'évaluation mais, depuis, le nombre a explosé. EvalPartners, sous les auspices de l'Organisation Internationale de Coopération en Évaluation (OICE), a recensé un total 158 associations ou réseaux, dont 135 au niveau national.

Par conséquent, l'évaluation devra s'adapter à une grande variété d'environnements opérationnels. Le processus devrait être accéléré. L'évaluation aujourd'hui n'est qu'une industrie naissante. À l'échelle mondiale, toutes les associations et tous les réseaux d'évaluation étudiés par EvalPartners ont totalisé 32 000 membres, en incluant le double comptage des membres qui appartiennent à plus d'une association. Cela représente moins d'un cinquième des membres d'une seule association d'auditeurs internes (par exemple, l'Institut des Auditeurs Internes compte 175 000 membres). Rien qu'aux États-Unis, il y a environ 1,2 million de comptables et d'auditeurs et auditrices.

Diversification

À mesure que l'évaluation franchira les frontières, elle devra s'étendre au-delà du secteur public, atteindre les entreprises privées et les fondations philanthropiques, ces nouveaux acteurs de la sphère publique, ainsi que les programmes de responsabilité sociale et environnementale des multinationales, et la croissance explosive des ONG. L'évaluation devra donc subir une révolution culturelle pour s'adapter aux nouvelles parties prenantes. Il faudra, en effet, faire preuve d'une plus grande agilité et les intégrer dans les pratiques.

Les évaluatrices et évaluateurs devraient se mettre au service de coalitions multisectorielles axées sur la réalisation d'objectifs mondiaux et régionaux précis, notamment en matière de santé publique, d'éducation et de protection de l'environnement. Alors que la pratique actuelle met l'accent sur les programmes au niveau des projets et des pays, l'accent devrait être mis progressivement sur l'évaluation de la capacité à générer des biens publics mondiaux et régionaux par des réseaux d'acteurs et d'actrices qui partagent les mêmes objectifs. Ces évaluations axées sur les réseaux devront mettre l'accent sur les mesures de l'impact collectif des coalitions et des partenariats (Liebenthal, Feinstein, et Ingram, 2004).

Numérisation

Les évaluateurs et évaluatrices doivent accepter le fait que nous vivons dans un monde « connecté ». Nous sommes inextricablement liés sur le plan social, financier et culturel au-delà les frontières, et le retour en arrière est impossible. Nous sommes en plein milieu d'une transformation mondiale de la société, silencieuse, progressive et irréversible.

Il faudra donc faire preuve de plus de rapidité et de réactivité dans la prestation des services d'évaluation. L'abandon du rythme statique des processus d'évaluation du secteur public sera accéléré par un autre courant sismique et omniprésent résultant de l'impact explosif des nouvelles technologies de l'information et des communications. Après l'époque de l'ordinateur central et du PC, il combine les énergies sociales déclenchées par le Web 2.0 et le potentiel analytique des *big data* associées à Web 3.0.

Le terme Web 2.0 évoque l'utilisation systématique de logiciels sociaux à toutes les étapes du processus d'évaluation. Il implique l'utilisation croissante d'applications informatiques et de *smartphones* pour faciliter les évaluations. Il s'appuie sur elles pour créer et publier le contenu de l'évaluation. Il permet de rapprocher les évaluateurs et évaluatrices, les gestionnaires de programme et les bénéficiaires et il offre de nouvelles façons de présenter les résultats. Au lieu de rapports longs et volumineux, les commanditaires se verront présenter des présentations virtuelles de textes brefs résumant des perspectives comparatives, des images vivantes et des vidéos, y compris des hyperliens qui permettront d'accéder aux contenus techniques et aux preuves à l'appui.

Le terme Web 3.0 est toujours contesté, mais toutes ses définitions font état du potentiel des moteurs de recherche qui compilent et passent au crible les torrents de données qui circulent actuellement sur le web mondial. D'énormes quantités de données numériques sont constamment créées. Des millions de capteurs sont incorporés dans les téléphones

portables, les distributeurs automatiques, les ordinateurs personnels, les tablettes, les véhicules de transport et les machines industrielles. Un volume phénoménal et croissant de données émerge lorsque les individus se déplacent, effectuent des transactions commerciales ou communiquent avec d'autres personnes par courriel, Skype, ou sur les réseaux sociaux. La révolution des *Big Data* va donc rendre l'évaluation plus agile et plus puissante.

Nouvelles méthodes et normes

L'évolution du contexte opérationnel et les nouvelles orientations politiques esquissées ci-dessus impliquent des ajustements dans les méthodes et les normes. Trois autres défis majeurs doivent également être relevés : le rééquipement pour répondre aux besoins des nouveaux commanditaires, la professionnalisation pour être compétitif sur le marché de l'évaluation et renforcer son indépendance, et la démocratisation pour promouvoir l'intérêt public.

Remise à neuf des outils

Les méthodes traditionnelles d'évaluation du développement reposent souvent sur des chaînes de résultats et des méthodes expérimentales qui évoquent des phénomènes sociaux linéaires, statiques et prévisibles. Elles sont mal adaptées à des contextes dynamiques caractérisés par la complexité, la non-linéarité et l'émergence. Pourtant, le marché de l'évaluation est dominé par des interventions qui sont vulnérables aux changements rapides de l'environnement opérationnel et aux pressions imprévisibles et parfois contradictoires d'un large éventail de parties prenantes.

Lorsque le changement est la seule constante, les valeurs éthiques, plus que les résultats prédéterminés, sont le moteur de l'évaluation. Par conséquent, dans le domaine de l'innovation sociale, le nouveau programme d'évaluation du développement devra être axé sur la valeur sociale et des boucles d'apprentissage plus courtes. La clarification des valeurs humaines permettra d'identifier des objectifs pertinents et de déterminer comment les atteindre, ce qui devrait aider à la prise de décisions idoines, et de définir quelles responsabilités distinctives et obligations réciproques devraient être adoptées.

Compte tenu d'un environnement opérationnel caractérisé par la volatilité et le changement, les essais contrôlés par assignation aléatoire ne seront plus considérés comme l'étalon d'or. Ils seront enfin reconnus comme n'étant qu'un des moyens d'évaluer l'inférence causale (Pearl et MacKenzie, 2018). L'évaluation d'impact partagera la vedette avec les analyses comparatives qualitatives, le suivi des processus, les réseaux bayésiens, etc. Des méthodes qualitatives seront adoptées, notamment les enquêtes, les focus group, les entretiens, les données de suivi, l'analyse comparative et les panels d'experts. La diversité méthodologique mettra les méthodes mixtes au service d'une évaluation sur mesure. Elle adoptera divers modèles d'évaluation – y compris des modèles réalistes, des études de cas, des modèles expérimentaux et quasi-expérimentaux – qui s'appuieront sur des données quantitatives et qualitatives.

Cela sera facilité par une méthode systémique axée sur les perspectives, les limites et les interrelations (Hummelbrunner et Reynolds, 2013). L'analyse des réseaux sociaux s'appuiera sur la sociologie, l'économie, les mathématiques et l'informatique pour cartographier, mesurer, évaluer et représenter les relations entre les individus et les groupes. Enfin, certains utilisateurs de l'évaluation exigeront des connaissances livrées juste à temps pour la prise de décision et rejetteront le rôle distant de l'évaluateur traditionnel du développement. Ce modèle réunira le suivi et l'évaluation, incorporera l'évaluation dans les processus de gestion et répondra à l'évolution des besoins conformément à l'approche de l'évaluation évolutive [*developmental evaluation*] (Patton, 2011).

L'impératif démocratique

Des forces puissantes ont façonné le monde de la post-vérité. L'évaluation ne constituera pas un contrepoids efficace aux intérêts égoïstes tant qu'elle reste éloignée du modèle démocratique. Selon House (2013), « la confiscation de l'évaluation par ses commanditaires est la plus grande menace que la communauté de l'évaluation ait connue depuis un certain temps. En fait, la crédibilité du domaine est menacée ». Sans éthique, l'institution d'évaluation est construite sur un terrain mouvant et ne peut pas survivre aux inévitables déluges. L'évaluation fondée sur l'équité deviendra la norme, et les évaluatrices et évaluateurs devront faire preuve de sens politique dans un contexte où la démocratie est menacée.

Les trois approches qui dominent l'évaluation aujourd'hui sont insuffisantes. La première, qui met l'accent sur la responsabilité et la conformité, examine comment les ressources publiques sont utilisées pour atteindre des objectifs qui sont presque toujours fixés par la structure du pouvoir en place. La deuxième est axée sur la recherche en sciences sociales. Elle met l'accent sur les évaluations axées sur l'attribution qui évoquent une approche scientifique objective, mais manque de conscience sociale. La troisième, l'évaluation axée sur l'utilisation, et qui s'apparente au conseil de gestion, a connu un succès remarquable, mais elle souffre souvent d'un manque d'indépendance (Patton, 2008).

Un autre modèle, l'évaluation démocratique, était plein de promesses lorsqu'il est apparu (Simons, 2010). Conçu par feu le professeur Barry MacDonald de l'Université d'East Anglia, il représente un service d'information à la communauté, qui confie aux évaluateurs et évaluatrices le rôle d'intermédiaire entre différents groupes. Il offre la confidentialité aux informateurs et informatrices et leur permet de contrôler l'information. Il ne tire pas de recommandations de ses conclusions (MacDonald, 1979). Cette méthode fonctionne bien dans les environnements où la rationalité de la communication prévaut et où le

discours éthique influence la prise de décision. Mais, compte tenu de sa position neutre de courtage, elle est mal adaptée aux contextes où elle est la plus nécessaire, dont notamment le monde de la post-vérité.

Pour promouvoir les intérêts des moins fortunés, House a affiné le modèle de MacDonald. Son modèle met l'accent sur trois principes : l'inclusion (travailler avec des groupes sous-représentés et impuissants), le dialogue (amener les parties prenantes à se comprendre) et la délibération (débat rationnel sur les questions, les valeurs et les conclusions). Dans cette incarnation révisée, « l'évaluateur ou l'évaluatrice n'est pas un-e spectateur/-trice passif/-ve, un-e facilitateur/-trice innocent-e ou un-e philosophe qui prend des décisions pour les autres, mais plutôt un-e professionnel-le consciencieux/-cieuse qui adhère à des principes mûrement réfléchis » (House et Howe, 1999). Il ne fait aucun doute que cette position militante est mieux adaptée aux environnements qui sont partiellement démocratiques.

Toutes les approches d'évaluation démocratique existantes sont entravées dans des contextes de post-vérité qui ne tolèrent pas la dissidence et qui sont donc étroitement contrôlés par les commanditaires de l'évaluation. Dans de telles situations, les progrès vers des idéaux démocratiques libéraux exigent un autre modèle : l'évaluation démocratique indépendante (Picciotto, 2015). Les évaluateurs et évaluatrices opérant selon ce modèle tireraient parti de l'influence croissante des parlementaires et de la société civile.

Ils et elles s'approprieraient les produits de l'évaluation et refuseraient les missions destinées à informer des responsables de l'intervention évaluée, rendant plutôt compte à une autorité suprême, comme un conseil d'administration ou un parlement, ou à des entités indépendantes de l'intervention, comme une ONG.

Professionnaliser l'évaluation

À l'heure actuelle, le grand public est mal informé de ce que représente l'évaluation. Les évaluateurs et évaluatrices sont régulièrement confondu-e-s avec les auditeurs/-trices et les chercheurs et chercheuses en sciences sociales. L'éducation et la formation en matière d'évaluation sont rares. La discipline n'est pas encore parvenue à un accord universel sur les principes directeurs, les directives éthiques et les compétences requises. Les évaluateurs et évaluatrices ne contrôlent pas l'accès à ce titre. Par conséquent, la qualité du travail d'évaluation est très variable. N'importe qui peut se présenter comme évaluateur ou évaluatrice.

En un mot, les évaluateurs et évaluatrices n'ont toujours pas le statut, le prestige et l'autonomie d'une profession. À l'avenir, les associations d'évaluation auront fort à faire pour accroître leur influence dans l'environnement opérationnel de la post-vérité. Elles devront plaider et œuvrer à la création d'un plus grand nombre d'évaluateurs et d'évaluatrices compétent-e-s en élargissant l'accès à des programmes d'éducation et de formation de grande qualité, tout en concevant et en mettant en œuvre des systèmes de certification d'évaluation, et en défendant la marque de l'évaluation en s'engageant dans des actions de plaidoyer auprès d'un plus grand nombre de client-e-s.

L'amélioration de la connectivité des évaluateurs et évaluatrices du développement au-delà des frontières ne sera possible que si des liens systématiques sont établis entre les groupes d'intérêt thématiques hébergés par les associations d'évaluation. À leur tour, les communautés épistémiques qui en résultent devront sortir de leurs cloisonnements disciplinaires, s'ouvrir aux professions connexes (comme l'administration publique, le conseil en gestion et l'audit) et contribuer à combler le fossé entre la théorie des sciences sociales et la recherche comportementale (Vaessen et Leeuw, 2010). [...]

Bibliographie

Chelimsky, Eleanor, et William R. Shadish. 1997. *Evaluation for the 21st century*. London and Thousand Oaks: Sage Publications.

Davis, Evan. 2017. *Post truth: Peak bullshit and what We can Do about It*. London: Little Brown.

House, Ernest R. 2013. « Evaluation's conflicted future ». P. 64 in *The future of evaluation in society: A tribute to Michael Scriven*, édité par S. I. Donaldson.

House, Ernest R., et Kenneth R. Howe. 1999. *Values in Evaluation and Social Research*. 1re éd. Thousand Oaks: Sage Publications.

Hummelbrunner, Richard, et Martin Reynolds. 2013. « Systems thinking, Learning and values in evaluation ». *Evaluation connections: The European Evaluation Society Newsletter* 9-10.

Liebenthal, Andrès, Osvaldo N. Feinstein et Gregory K. Ingram. 2004. « Evaluation and development: The partnership dimension ». in *World Bank series on evaluation and development*. New Brunswick and London: Transaction Publishers.

MacDonald, Barry. 1979. *Democracy and evaluation*. Mimeo, Centre for applied research in evaluation. University of Anglia.

Patton, Michael Q. 2008. *Utilization-focused evaluation*. 4e éd. Los Angeles: Sage Publications.

Patton, Michael Q. 2011. *Developmental evaluation, applying complexity concepts to enhance innovation and us*. New York and London: The Guilford Press.

Pearl, Judea, et Dana MacKenzie. 2018. *The book of why: The New science of cause and effect*. UK: Allen Lane, Penguin Random House.

Picciotto, Robert. 2015. « Democratic evaluation for the 21st century ». *Evaluation* 21(2) : 150-66.

Simons, Helen. 2010. *Democratic evaluation: Theory and practice*. Virtual Evaluation Conference.

Vaessen, Jos, et Frans L. Leeuw. 2010. *Mind the gap: Perspectives on policy evaluation and the social sciences*. 1re éd. New Brunswick: Transaction Publishers.

Le regard de Nathalie Mons

NATHALIE MONS

Je ne me considère pas comme une professionnelle de l'évaluation. Ce sont mes activités de recherche qui m'ont conduite à m'ouvrir sur ce champ d'expertise. Sociologue spécialisée dans les politiques éducatives, mes recherches se sont dès le début orientées vers l'analyse des effets des réformes dans ce champ spécifique de l'action publique. Je maniais alors les analyses statistiques quantitatives et la comparaison internationale, notamment les grandes enquêtes de l'OCDE comme PISA pour mettre en évidence les pratiques et politiques éducatives qui pouvaient sembler les plus efficaces pédagogiquement.

Dès le début l'utilité sociale et citoyenne a été centrale dans mon travail. L'université était pour moi une seconde carrière après un parcours dans le secteur privé, dans lequel en tant que cadre vous devez en permanence rendre des comptes sur les investissements que vous avez sollicités. Il m'a toujours semblé que les responsables politico-administratifs qui mobilisent des deniers publics, c'est-à-dire l'argent commun, devraient plus encore que les acteurs du secteur privé s'astreindre à cette logique éthique de reddition des comptes. À mon arrivée dans le secteur public, j'ai été étonnée de voir la place marginale occupée par l'évaluation là où éthiquement elle me paraissait devoir être centrale.

Par la suite, j'ai affiné mes méthodes de recherche : j'ai compris que pour réellement évaluer les effets des politiques, des analyses de terrain plus qualitatives permettant de comprendre les processus par lesquels se créent les phénomènes étudiés devaient compléter la photographie centrale d'un phénomène qu'offre toute analyse quantitative. L'évaluation doit analyser ce qui se passe, évaluer, mesurer un phénomène, bien sûr, mais aussi développer une démarche de recherche qui permette de comprendre les raisons souvent complexes des phénomènes étudiés, sinon son utilité et donc ses usages ne peuvent être que limités. C'est

cette approche qui m'a fait accepter en 2013 la présidence du Conseil national d'évaluation du système scolaire (Cnesco) chargé en France de conduire une évaluation indépendante de l'école française puis une chaire au CNAM sur l'évaluation des politiques publiques d'éducation.

Par ailleurs, comprenant rapidement que le métier d'évaluateur ou d'évaluatrice est un métier dangereux pour la personne concernée, mais aussi pour les personnes évaluées (car l'évaluation peut orienter très rapidement les pratiques professionnelles), j'ai souhaité également développer des recherches sociologiques sur les usages et réceptions des évaluations par les décideurs politiques et les acteurs de terrain : par exemple, comment un instrument comme l'enquête de l'OCDE PISA qui évalue les élèves de 15 ans et fait la une des grands médias nationaux dans les pays de l'OCDE est-elle utilisée, voire manipulée par les politiques pour imposer la pertinence de certaines réformes? Ou comment se comportent les enseignants et les élèves face aux grandes enquêtes nationales et internationales? Les individus sont stratèges et apportent toujours aux décideurs et décideuses les résultats qu'ils et elles souhaitent si ces dernier-e-s développent des évaluations à forts enjeux.

L'évaluation d'une pratique ou d'une politique publique n'est pas un acte anodin. L'évaluation peut avoir des effets délétères, tout-e évaluateur ou évaluatrice devrait faire un travail réflexif sur sa démarche, ses outils et sa communication.

Plus largement, le fait d'être non pas une praticienne de l'évaluation, mais une chercheuse spécialisée dans l'élaboration et les effets des politiques publiques m'a permis très tôt de comprendre que la place de l'évaluation est, de fait, et doit, éthiquement, être limitée dans le processus décisionnaire politique sinon nous ne sommes plus en démocratie. Le savant doit respecter l'autonomie du politique. Ce parcours de recherche m'empêche d'avoir une vision mécaniste et donc naïve sur la thématique de l'usage des évaluations.

Le recensement d'articles scientifiques effectué dans ce chapitre est passionnant, car les textes sur les usages des évaluations ne sont pas légion. Il existe une littérature très riche sur les méthodologies de l'évaluation, mais le champ de ses usages a été moins exploré. L'évaluateur ou l'évaluatrice présuppose trop souvent que son travail, s'il est de qualité, sera utilisé, d'où cette très forte centration dans les recherches sur les dimensions méthodologiques et non politiques ou d'usage de l'évaluation.

Or, le premier résultat convergent de ces articles est que l'usage des évaluations dans le design de futures politiques publiques reste encore marginal. On trouve ce résultat dans les tous les pays développés et sur tous les secteurs de l'action publique. Mais en ouvrant la focale de la définition des usages, comme le font de nombreux textes du chapitre, on peut alors étudier des utilisations largement plus présentes, notamment par le biais d'influences indirectes.

Plutôt qu'un regard technique sur l'évaluation, quand on s'intéresse à ses usages, il faut embrasser plus largement le processus décisionnaire de politiques publiques et se demander comme le font certain-e-s auteurs et autrices présenté-e-s dans le chapitre : quels sont les ingrédients techniques, politiques, démocratiques de la décision publique? Dès lors que l'on réfléchit ainsi sur le lien entre démocratie et évaluation, il apparaît que la place de l'évaluation dans la décision publique et donc ses usages ne peut être qu'à la marge. L'équipe d'évaluation éclaire ou informe le processus décisionnaire de politique publique au mieux.

Dans un programme de recherche, nous avons cherché à savoir si, dans les pays ou dans les gouvernements qui se présentaient comme des adeptes de l'approche *evidence based policy* (politique publique fondée sur les preuves), nous pouvions retrouver des exemples concrets de cette démarche dans les réformes qu'ils mettaient en place sur le terrain. Nous n'en n'avons jamais trouvé, même dans une nation comme l'Angleterre qui a développé à marche forcée des *What Works Centres* dans de nombreux secteurs de l'action publique. L'approche *evidence-based policy* reste un modèle théorique. Le politique peut à certaines époques, dans certains

pays développer un discours sur la légitimité de politiques publiques qui serait fondée sur des évaluations ou des recherches, mais heureusement les politiques publiques sont des processus démocratiques beaucoup plus complexes. Les politiques ne sont jamais les transpositions simples de résultats de recherche ou d'évaluation.

Avant d'être technique et méthodologique, la réflexion de l'évaluateur ou de l'évaluatrice qui souhaite voir ses résultats partagés par une communauté politique ou de praticien-ne-s d'un secteur d'activité doit être axiologique et éthique : quelle doit être ma place dans cette communauté que j'ambitionne d'éclairer de mes résultats, à quelles conditions une attention portée à mes résultats peut-elle exister? De quelle légitimité disposé-je par rapport à l'expertise et à la légitimité des décideurs/décideuses ou acteurs/actrices de terrain? Lorsque l'on réfléchit ainsi, je pense que rapidement des postures d'humilité, de respect de l'autonomie politique et professionnelle des évalué-e-s et de curiosité pour leur expertise s'imposent. L'évaluateur ou l'évaluatrice qui veut avoir une chance de rentrer en dialogue avec une communauté professionnelle qu'il ou elle évalue ne doit pas être dans une démarche d'imposition de ses résultats, une approche verticale, car il ou elle n'en a la légitimité ni politique ni même parfois technique.

Le métier d'évaluateur ou d'évaluatrice est ingrat, car cette posture d'humilité doit se doubler d'un travail intellectuellement très robuste sans lequel l'évaluation n'aura pas non plus de légitimité. Durant la mission de préfiguration du Conseil national d'évaluation du système scolaire, j'ai réalisé plus d'une cinquantaine d'interviews de décideurs et décideuses politiques, praticien-ne-s (enseignant-e-s, chef-fe-s d'établissement), parents… pour essayer de comprendre ce qu'ils et elles attendaient de ce nouvel organisme. Contrairement à ce que pensent parfois les évaluateurs/-trices, qui s'enferment dans une forme d'arrogance et verticalité du/de la savant-e, les professionnels de terrain ont un regard très aiguisé sur la qualité des évaluations. Seules sont légitimes à leurs yeux des évaluations méthodologiquement de qualité, mais aussi des évaluations qui ne s'enferment pas dans une chapelle de pensée. Les

évaluations qui n'observent les phénomènes qu'à travers un prisme disciplinaire unique ou un objectif exclusif (par exemple budgétaire) ne sont pas légitimes, car elles conduisent à une vision biaisée et parcellaire des phénomènes à étudier, elles ne sont pas perçues comme rendant justice au travail des professionnel-le-s de terrain ou des responsables politiques.

Pour être partagée à terme, une évaluation doit être considérée comme légitime. Elle ne peut être que complexe, faire intervenir des entrées disciplinaires multiples, des niveaux territoriaux variés, croiser des méthodes quantitatives et qualitatives. Par exemple, au Cnesco quand nous avons travaillé sur les politiques de redoublement, sociologues, didacticien-ne-s, économistes et psychologues ont analysé ce phénomène très français pour en comprendre les effets psychosociaux, didactiques et économiques tous négatifs, mais aussi les effets sociologiques qui permettaient de comprendre pourquoi les parents et les professeur-e-s eux-mêmes et elles-mêmes restaient très attaché-e-s à cette pratique dont quatre décennies de recherche avaient pourtant montré la totale inefficacité en termes d'apprentissage et les effets psychologiques désastreux sur les élèves du primaire. On voit bien là la complémentarité des entrées disciplinaires qui permet d'analyser de façon holistique un phénomène. Les évaluations qui prennent la forme de tableaux de bord avec quelques indicateurs statistiques rustiques ont peu de chance de retenir l'attention des évalué-e-s.

Scientifique, acceptant le jeu de la complexité, l'évaluation doit aussi être participative. Il faut enrôler les professionnel-le-s avec lesquel-le-s on souhaite à terme partager des résultats dès le début de l'opération dans un dispositif qui, cependant, ne leur donne pas la main sur les résultats de recherche. Au Cnesco, les professionnel-le-s de terrain n'ont pas leur mot à dire sur les contenus des rapports que nous commandons à des chercheurs et chercheuses. En revanche, en amont ils et elles nourrissent notre réflexion sur les questions de recherche que nous poserons aux évaluateurs et évaluatrices, et surtout participent à nos côtés, par exemple comme juré-e-s de nos conférences de consensus, pour écrire

les recommandations que nous tirons des résultats de nos évaluations. Après la phase de production des résultats des évaluations, la production de recommandations doit retourner vers la réalité de terrain. Les communautés de praticien-ne-s ou les responsables politiques ne pardonnent pas aux équipes d'évaluation le hors-sol. L'équipe d'évaluation doit accepter qu'à la phase d'élaboration des recommandations l'expertise de l'évalué-e puisse être largement supérieure à la sienne.

Le futur de l'évaluation se jouera dans la participation citoyenne. Soit l'évaluation a la capacité de se transformer en outil démocratique capable de revitaliser notre démocratie participative par une action plus directe des citoyen-ne-s sur les politiques publiques avec ainsi des usages réels, soit son impact stagnera. Cela nécessite en amont un rééquilibrage de la séparation et de l'équilibre des pouvoirs entre le Parlement et l'exécutif, car ce dernier n'acceptera jamais aisément de se voir imposer un processus de jugement de la qualité de son travail. D'ailleurs, les grands pays d'évaluation des politiques publiques ont confié ce pouvoir au Parlement; ou plutôt le Parlement (je pense au Congrès américain) s'en est emparé. L'évaluation et ses usages supposent toujours un rapport de force violent.

II. QUI ÉVALUE?

Introduction : qui évalue et comment?

THOMAS DELAHAIS, AGATHE DEVAUX-SPATARAKIS, ANNE REVILLARD ET VALÉRY RIDDE

À partir du moment où l'évaluation est devenue un champ à part (partie Science), deux questions se sont posées : Qui évalue? Quel est le rôle de celles et ceux qui évaluent?

Qui évalue?

Pendant longtemps, les évaluations ont été le fait de chercheurs et de chercheuses, considérant l'évaluation comme une activité de recherche parmi d'autres, et appliquant leurs méthodes à des terrains d'investigation nouveaux (voir partie Science).

À partir des années 1970, l'activité évaluative s'élargit au-delà du monde académique. Aux États-Unis d'abord, puis dans le monde entier, fonctionnaires, consultant-e-s, personnels associatifs, etc. sont amené-e-s à évaluer. Concomitamment, les figures de l'évaluateur et de l'évaluatrice s'autonomisent. Les administrations créent des postes de chargé-e-s d'évaluation internes, dont le rôle et les missions sont proches, mais distincts de celles des évaluateurs et des évaluatrices externes[1]. Progressivement, être évaluateur ou évaluatrice devient un métier nécessitant des compétences et des approches particulières, et

1. Arnold Love (Love, 1991) place les évaluateurs/-trices internes au service du management des organisations, avec un rôle essentiel en termes de planification de l'évaluation sur la durée (là où les équipes externes ont un rôle plus ponctuel), mais aussi de résolution de problèmes et d'aide à la décision. Ceux-ci et celles-ci savent quelles informations sont nécessaires à leurs collègues pour faire des choix, où les trouver si elles existent et le cas échéant comment les obtenir – y compris en faisant appel à une évaluation externe.

on voit apparaître des sociétés professionnelles, des cursus de formation initiale ou continue, voire des certifications spécifiques comme au Canada (Gauthier, 2020).

Sagesse pratique et posture évaluative

Aujourd'hui, les deux visions cohabitent. D'une part, celle de l'évaluation comme une activité parmi d'autres, dans laquelle celles et ceux qui font les évaluations font appel aux principes, à l'expertise, aux savoir-faire propres à leur profession. D'autre part, celle de l'évaluation comme métier spécifique.

Une question se pose toutefois : dans une optique de professionnalisation, la pratique évaluative peut-elle se limiter à une liste de compétences et d'aptitudes particulières? Ce qui caractérise l'évaluateur et l'évaluatrice n'est pas tant l'emploi de méthodes ou leur capacité à animer des ateliers, que leur posture évaluative (*evaluative thinking*). Celle-ci est d'abord « une application de la pensée critique dans le contexte de l'action publique. Elle se caractérise par une attitude curieuse, et la volonté de fonder sa démarche sur des éléments de preuve. Rentrer dans une posture évaluative, c'est identifier des hypothèses et poser les bonnes questions de façon à mieux comprendre, à travers une réflexion approfondie et la prise de recul, les phénomènes observés, et ainsi éclairer l'action et la prise de décision » (Buckley *et al.*, 2014, notre traduction). Notons ici qu'il existe un texte en français sur ce sujet (Archibald et Moussavou, 2016).

De leur côté, Thomas Schwandt et Ernest House à sa suite ont emprunté à Aristote le concept de sagesse pratique (*phronesis*) pour rendre compte de ce qui, au-delà des compétences, singularise les évaluateurs et les évaluatrices. Comment mener une évaluation à bien, la rendre utile, faire en sorte qu'elle contribue au bien commun? Certainement pas en employant des solutions toutes faites.

> La sagesse pratique s'oppose notamment à toute application mécanique de règles trop abstraites [...], de procédures formalisées, de savoirs scientifiques ou de routines. Elle requiert donc notamment une attention particulière aux caractéristiques concrètes de ces cas ou situations, en sorte [d'en] saisir la complexité et la singularité. Mais l'observation et l'analyse approfondies ne permettent pas de lever toute incertitude. C'est pourquoi la sagesse pratique est aussi conjecturale. Pour pouvoir agir, l'homme prudent doit parfois accepter de faire des paris. (Champy, 2017)

Arnold Love le résume ainsi : la sagesse pratique est ce qui permet de « faire les bons choix, au bon moment, pour les bonnes raisons » (Love, 2018). Un livre en français explore largement cette idée (Hurteau, Bourgeois, et Houle, 2018). Mais la sagesse pratique ne s'acquiert-elle que par l'expérience et la réflexivité? Ne devrait-elle pas, au contraire, être au cœur de l'enseignement de l'évaluation?

Quelle responsabilité pour les évaluateurs et les évaluatrices?

Les évaluateurs et les évaluatrices des années 1950 et 1960 pensaient qu'il était possible de résoudre les problèmes sociaux en découvrant « la vérité » sur ce qui fonctionne (ou dysfonctionne), et se voyaient avant tout comme des garant-e-s de la méthodologie. Il s'agissait de choisir la bonne approche, de mener au mieux les opérations techniques et de retranscrire fidèlement les résultats de la démarche pour tendre à l'objectivité scientifique (voir partie Science). Il revenait ensuite à la sphère de la décision de retenir ce qui leur paraissait utile à l'amélioration des politiques.

Dès les années 70 (voir partie Utilité), l'idée qu'il suffit d'établir la vérité pour que les évaluations soient utilisées fait toutefois long feu; il en va de même de la capacité des évaluateurs et des évaluatrices à déterminer seul-e-s ce qui est bon pour la société. La question qui est alors posée est d'abord celle de leur responsabilité dans la conduite de l'évaluation. Cela implique notamment de : i) concevoir et mettre en œuvre des évaluations robustes, en fonction du moment où une décision doit être prise (ce qui entraîne des contraintes fortes de données, de budget et de temps disponible); ii) savoir repérer ce que les parties prenantes cherchent à évaluer et les aider à l'établir précisément; iii) enfin, cerner les conditions d'utilisation de l'évaluation (par qui, pour quoi, comment).

Bien entendu, la façon dont les évaluateurs et les évaluatrices ordonnent ces trois dimensions (méthode, valeur, utilité) est aussi l'objet de controverses multiples, qui témoignent de visions distinctes de leur métier. Mertens et Wilson (2012) ont regroupé les différents points de vue en quatre « paradigmes » de l'évaluateur/-trice : positiviste (privilégiant les méthodes et l'apport de connaissances et de concepts); pragmatique (recherchant d'abord l'utilité des travaux et la contribution à la prise de décision); constructiviste (se voyant comme maïeuticien-ne et médiateur/-trice, permettant à chacun-e d'expliciter ses propres valeurs au service des échanges entre parties prenantes); transformationnel (visant avant tout à changer le monde au service du bien commun, aux côtés des groupes dominés). Dans les trois parties ci-dessous, nous revenons sur les controverses qui ont ainsi agité la communauté évaluative au sujet du rôle de l'évaluateur/-trice.

Fondements : vers un métier évaluatif

Dès les années 1960, Edward Suchman (voir partie Science) met en évidence une différence majeure entre les chercheurs/-euses dans leurs activités d'évaluation et les évaluateurs/-trices. En effet, le plus souvent,

ces dernier-e-s ne maîtrisent pas les conditions dans lesquelles va s'effectuer l'évaluation, laquelle doit pourtant respecter des critères de rigueur et de crédibilité. Or, les équipes d'évaluation sont généralement sollicitées à un moment donné du déroulé d'une intervention. Elles ne sont pas présentes lors de sa conception, ne peuvent agir sur la mise en place du système de suivi et d'évaluation, ni s'assurer de la qualité ou de la pertinence des données produites au fur et à mesure. La commande évaluative arrive tardivement et le budget qui lui est consacré est souvent réduit (en particulier eu égard aux enjeux relatifs aux données ou au temps disponibles évoqués plus haut).

C'est cet aspect consubstantiel du métier qu'abordent Michael Bamberger et al. dans un texte de 2004 (**texte 1**). Les auteurs et autrices y proposent une certaine approche, celle de l'évaluation « en situation réelle » (*shoestring evaluation*), une démarche heuristique pour résoudre la quadrature du cercle dans laquelle se trouve l'équipe d'évaluation. Un des aspects essentiels qui y est présenté, est qu'il est rare de pouvoir partir de la méthode idéale et de l'adapter, car alors, les arrangements effectués risqueraient de compromettre l'évaluation tout entière. Il est donc nécessaire d'élaborer à chaque fois une stratégie ad hoc, répondant aux besoins effectifs d'information du ou de la commanditaire, et réalisable dans les conditions (de temps, de budget) imparties.

Au-delà des questions de méthode, les évaluateurs et les évaluatrices construisent aussi progressivement une vision de ce que « bien faire » son métier signifie. À partir des cinq « principes directeurs » de l'Association américaine d'évaluation (AEA) : l'investigation systématique, la compétence, l'intégrité/l'honnêteté, le respect pour les personnes et les responsabilités vis-à-vis du bien commun, Michael Morris (**texte 2**) rend compte des débats relatifs à l'éthique qui ont agité la communauté évaluative à travers deux de ses journaux, *Evaluation Practice* et *American Journal of Evaluation*. Comment réagir aux pressions politiques, qu'elles prennent la forme de contraintes, ou d'une connivence tout au long de la démarche d'évaluation? L'équipe d'évaluation est-elle à même d'éviter le mauvais usage qui peut être fait de son travail? Comment gérer par

exemple le dilemme entre la possibilité d'avoir un meilleur taux de réponse à une enquête et s'assurer du consentement explicite et éclairé de chacun-e? Entre l'utilisation d'une méthode innovante et le risque de divulguer des données privées? Pour Morris, la capacité à traiter ces enjeux éthiques en situation devrait faire partie des compétences nécessaires à la pratique évaluative.

Controverses : quel rôle pour l'évaluateur ou l'évaluatrice?

Dans les années 1990 et 2000 s'est largement posée la question du rôle des évaluateurs et des évaluatrices dans l'action publique. Doit-on se satisfaire de ce rôle « canonique » d'une évaluation extérieure qui se veut détachée des considérations politiques, et vise d'abord à répondre à la demande des commanditaires – ou bien doit-on au contraire, embrasser pleinement la dimension politique de l'évaluation?

Les débats se cristallisent sur la question de savoir si les évaluateurs et les évaluatrices doivent prendre la défense d'une des parties prenantes au détriment d'une autre. Ernest House et Kenneth Howe (**texte 3**) rejettent l'idée que les équipes d'évaluation devraient soutenir des solutions ou des points de vue particuliers, au motif que cela ruinerait leur crédibilité. Pour autant, il relève de leur responsabilité que les utilisateurs et utilisatrices des évaluations aient accès à toutes les données pertinentes sur un sujet, en particulier lorsque celles-ci sont contradictoires, et que tous les éléments de preuve présentés par chaque partie soient rigoureusement contre-expertisés. De plus, il ne suffit pas d'exposer les points de vue de chacun-e : encore faut-il créer, entre les parties prenantes, les conditions du dialogue et de la délibération tout au long du processus évaluatif. Les auteurs nous avertissent : si cette « évaluation démocratique délibérative » constitue un idéal plus qu'une méthode, il revient aux évaluateurs et aux évaluatrices de faire de leur mieux pour l'atteindre.

Faut-il aller plus loin? Est-ce à l'équipe d'évaluation de prendre la décision d'orienter son travail en défense de certaines parties prenantes, notamment les plus marginalisées? En se plaçant au sein du paradigme transformationnel, Donna Mertens répond par l'affirmative **(texte 4)**. Pour elle, House et Howe ont certes proposé une approche incluant les parties prenantes les moins bien représentées, mais sans désigner l'origine des discriminations qu'elles subissent, telles que le racisme ou le sexisme. Les évaluateurs et les évaluatrices ont le devoir de nommer ces fléaux, mais aussi de reconnaître que leurs pratiques et la communauté qu'ils forment font partie du problème. Autrement dit, comment faire en sorte que les groupes traditionnellement discriminés soient représentés parmi les évaluateurs et les évaluatrices (Mertens, 2002)? Donna Mertens décrit ici ce que signifie la posture transformationnelle en repartant de son parcours personnel, notamment auprès des personnes sourdes et malentendantes.

Perspectives : l'évaluation en défense du bien commun

Sandra Mathison enfonce le clou de la remise en cause collective. Dans son introduction aux débats des Journées de la société australienne d'évaluation en 2017, elle pose la question : « L'évaluation contribue-t-elle au bien commun? » **(texte 5)**. Les évaluateurs et les évaluatrices sont pénétré-e-s de l'utilité de leur profession pour la société. En rendant compte des conséquences des interventions sur leurs destinataires, ne permettent-ils et elles pas de les améliorer au bénéfice de tou-te-s?

Or, pour l'autrice, ce n'est pas dans l'atteinte des résultats que se situe l'enjeu fondamental du métier évaluatif, mais plutôt dans la mise en lumière de l'idéologie sous-tendant les interventions évaluées, des choix qui sont faits des problèmes à résoudre, ou encore dans l'interconnexion entre les principaux maux (environnementaux, économiques, sociaux) qui

affectent la société. Or les évaluateurs et les évaluatrices se contentent de répondre aux questions qui leur sont posées par leurs commanditaires. Pour Mathison, l'évaluation, telle qu'elle se fait aujourd'hui, manque ainsi fondamentalement d'indépendance : elle se glorifie de « dire la vérité au pouvoir », qui la connaît déjà ou s'en désintéresse, sans faire l'effort d'apporter des « vérités qui dérangent », ou de battre en brèche les évidences de l'idéologie libérale – par exemple, la soi-disant responsabilité des pauvres dans leur situation.

Pour Mathison, la contribution de l'évaluation au bien commun ne va pas de soi : charge aux évaluateurs et aux évaluatrices d'en faire une réalité.

Bibliographie

Archibald, Thomas, et Laurent Ogoueli Moussavou. 2016. « La "pensée évaluative" : une activité mystérieuse et quotidienne ». *Éducation permanente* (208) : 33-40.

Buckley, Jane, Thomas Archibald, Monica Hargraves et William M. Trochim. 2014. « Defining and Teaching Evaluative Thinking ». *American Journal of Evaluation* 36(3) : 375-388. doi : https://doi.org/10.1177/1098214015581706.

Champy, Florent. 2017. « Décrire des activités prudentielles pour aider à les réhabiliter? Enjeux théoriques et pratiques d'enquêtes qualitatives sur la prise en charge de malades précaires dans les permanences d'accès aux soins de santé en France ». *Recherches qualitatives* 36(2) : 153-172.

Gauthier, Benoît. 2020. « Une analyse engagée de la professionnalisation des pratiques d'évaluation ». *The Canadian Journal of Program Evaluation* 35(1). doi : https://doi.org/10.3138/cjpe.69364.

Hurteau, Marthe, Isabelle Bourgeois et Sylvain Houle. 2018. *L'évaluation de programme axée sur la rencontre des acteurs : une sagesse pratique*. Québec : Presses de l'Université de Québec.

Love, Arnold J. 1991. *Internal evaluation: building organizations from within*. Newbury Park: Sage Publications.

Love, Arnold J. 2018. « 5. De la sagesse pratique à une pratique empreinte de sagesse ». in *L'évaluation axée sur la rencontre des acteurs une sagesse pratique*. M. Hurteau, I. Bourgeois, S. Houlle (éds). Québec : PUQ.

Mertens, Donna M. 2002. « The evaluator's role in the transformative context ». in *Exploring evaluator role and identity*, Ryan K.E. & Schwandt T. eds. IAP, p. 103-18.

Mertens, Donna M., et Amy T. Wilson. 2012. *Program Evaluation Theory and Practice*: A *Comprehensive Guide*. Guilford Press.

1. L'évaluation en situation réelle : concevoir des évaluations d'impact sous contraintes de budget, de temps et de données

MICHAEL BAMBERGER, JIM RUGH, MARY CHURCH ET LUCIA FORT

[Traduit[1] de : Bamberger, Michael, Jim Rugh, Mary Church et Lucia Fort. 2004. « Shoestring Evaluation: Designing Impact Evaluations under Budget, Time and Data Constraints », *American Journal of Evaluation*, 25(1) : 6-9. (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Le présent document traite de l'approche de l'évaluation en situation réelle (*shoestring evaluation*) qui est en cours d'élaboration afin d'aider les évaluateurs et les évaluatrices à effectuer des évaluations aussi fiables que possible sur le plan méthodologique lorsqu'ils et elles travaillent avec des contraintes budgétaires et temporelles et avec des limites quant au type de données auxquelles elles et ils ont accès. Cette approche est utilisée dans deux scénarios principaux. Le premier se produit lorsque l'évaluateur ou l'évaluatrice n'est pas appelé-e avant que le projet ou le programme ne soit opérationnel depuis un certain temps et qu'en général aucune donnée de base n'a été recueillie sur la population du projet ou

1. NdT : L'approche d'évaluation en situation réelle a également fait l'objet d'une présentation en français dans : Bamberger, M. et J. Rugh. 2012. "Une stratégie pour composer avec les contraintes inhérentes à la pratique", in V. Ridde et C. Dagenais. *Approches et pratiques en évaluation de programme*, Presses de l'Université de Montréal, p. 161-177.

sur un groupe témoin. Les gestionnaires, les décideurs et décideuses ainsi que les organismes de financement ne commencent souvent à se préoccuper de l'évaluation des impacts que lorsque le moment est venu de prendre des décisions sur le financement futur. L'évaluateur ou l'évaluatrice sera ainsi souvent obligé-e de travailler dans un calendrier inadapté et avec un budget limité. Le deuxième scénario se présente lorsque l'évaluatrice ou l'évaluateur est appelé-e au début du projet, mais que pour des raisons budgétaires, politiques, logistiques ou méthodologiques, il n'a pas été possible de recueillir des données de référence sur un groupe témoin, voire sur la population du projet elle-même, en utilisant des méthodologies comparables à une évaluation ultérieure.

Pour répondre à la demande croissante d'évaluations faisant face à ces contraintes budgétaires et de temps, un certain nombre de méthodes d'évaluation rapides et économiques ont été mises au point. Malheureusement, afin d'obtenir des résultats d'évaluation dans les délais et dans les limites du budget, bon nombre des principes de base d'une bonne conception de l'évaluation, comme l'échantillonnage aléatoire, l'explicitation de la théorie du programme, la mise au point d'instruments adaptés, le contrôle des biais des chercheurs et des chercheuses et le contrôle général de la qualité, peuvent se trouver compromis. L'approche de l'évaluation en situation réelle fournit des outils pour travailler dans les limites du budget, du temps et des données, tout en fournissant un cadre pour déterminer les menaces à la validité ou à la pertinence des conclusions de l'évaluation, ainsi que des lignes directrices pour faire face aux différentes menaces une fois qu'elles ont été identifiées.

À l'origine, l'approche a été élaborée pour aider les évaluateurs et les évaluatrices travaillant dans les pays en développement, où les contraintes budgétaires, temporelles et de données sont souvent les plus sévères. Toutefois, les commentaires de collègues travaillant aux États-Unis et dans d'autres pays industrialisés donnent à penser que l'approche

pourrait être plus largement applicable. Néanmoins, toutes les études de cas présentées dans le document sont tirées de l'expérience des auteurs et des autrices dans les pays en développement.

La plupart des outils et des méthodes utilisés dans l'approche seront familiers aux évaluateurs et aux évaluatrices expérimenté-e-s. Ce qui est nouveau, c'est la façon dont les outils sont combinés en une stratégie en six étapes afin d'assurer la meilleure qualité d'évaluation dans le cadre du budget, du temps et des contraintes en matière de données qui influent sur l'évaluation. Par conséquent, la plupart des méthodes de collecte et d'analyse des données ne sont mentionnées que brièvement. Nous discutons cependant de certaines des méthodes moins familières, telles que l'utilisation de rappels et d'autres méthodes pour reconstruire les données de l'étude de référence (*baseline* study) et des groupes témoins, les forces et les faiblesses de différents modèles quasi expérimentaux pour traiter les trois ensembles de contraintes, l'élaboration d'un cadre intégré pour évaluer la validité et l'adéquation des plans d'évaluation multiméthodes et les stratégies pour faire face aux différentes menaces à la validité et à l'adéquation. L'objectif est de mener des évaluations crédibles et adaptées aux besoins des principales parties prenantes, compte tenu des conditions dans lesquelles ces évaluations doivent être entreprises.

Scénarios d'évaluation en situation réelle : contraintes de temps, de données et de budget typiques auxquelles l'évaluateur ou l'évaluatrice doit faire face

Le tableau 1 décrit les scénarios d'évaluation typiques dans lesquels l'évaluateur ou l'évaluatrice est confronté-e à des contraintes liées au budget, au temps et aux données. Dans certains cas, l'évaluateur ou l'évaluatrice est confronté-e à une seule contrainte lorsque, par exemple, le budget est limité mais que l'évaluateur/-trice n'est pas confronté-e

à des contraintes de temps excessives, alors que dans d'autres cas, la principale contrainte est le temps. Dans d'autres cas, l'évaluation peut être planifiée avant le début du projet et il existe un budget adéquat, mais l'évaluateur ou l'évaluatrice est informé-e que, pour des raisons politiques ou éthiques, il ne sera pas possible de recueillir des données sur un groupe témoin. Beaucoup d'évaluateurs et d'évaluatrices malchanceux-ses se trouvent simultanément confronté-e-s à deux ou trois contraintes! Les paragraphes qui suivent traitent de certains des problèmes les plus courants rencontrés dans le cadre de chacune de ces contraintes.

Tableau 1. Scénarios d'évaluation en situation réelle

Les contraintes à l'égard desquelles l'évaluation doit être effectuée			**Scénarios typiques**
Temps	Budget	Données	
X			L'évaluateur/-trice est appelé-e à la fin du projet et on lui dit que l'évaluation doit être terminée à une certaine date afin qu'elle puisse être utilisée dans un processus décisionnel ou contribuer à l'élaboration d'un rapport. Le budget peut être adéquat, mais il peut être difficile de recueillir ou d'analyser les données d'enquête dans les délais prévus.
	X		L'évaluation n'est dotée que d'un petit budget, mais il n'y a pas nécessairement de contraintes de temps excessives. Toutefois, il sera difficile de recueillir des données d'enquête par sondage en raison du budget limité.
		X	L'évaluateur/-trice n'est pas appelé-e tant que le projet n'est pas bien avancé. Par conséquent, aucune étude de référence n'a été effectuée ni sur la population du projet ni sur un groupe témoin. L'évaluation a une portée suffisante, soit pour analyser les données existantes des enquêtes auprès des ménages, soit pour recueillir des données supplémentaires. Dans certains cas, l'impact prévu du projet peut également concerner des changements dans des domaines sensibles tels que la violence familiale, les conflits communautaires, l'autonomisation des femmes, les styles de *leadership* communautaire ou la corruption sur lesquels il est difficile de recueillir des données fiables – même lorsque le temps et le budget ne sont pas des contraintes.
X	X		L'évaluateur/-trice doit fonctionner avec des contraintes de temps et un budget limité. Des données d'enquêtes secondaires peuvent être disponibles, mais il y a peu de temps ou de ressources pour les analyser.
X		X	L'évaluateur/-trice a peu de temps et n'a pas accès aux données de l'étude de référence ou sur le groupe témoin. Des fonds sont disponibles pour recueillir des données supplémentaires, mais la conception de l'enquête est limitée par les délais serrés.
	X	X	L'évaluateur/-trice est appelé-e tardivement et n'a pas accès aux données de l'étude de référence ou sur le groupe témoin. Le budget est limité, mais le temps n'est pas une contrainte.
X	X	X	L'évaluateur/-trice est appelé tard, reçoit un budget limité, n'a pas d'accès aux données de l'étude de référence et aucun groupe témoin n'a été identifié.

Contraintes temporelles

La contrainte de temps la plus fréquente apparaît quand l'évaluateur ou l'évaluatrice n'est appelé-e que lorsque le projet est déjà bien avancé et que l'évaluation doit être effectuée dans un délai beaucoup plus court que ce qu'il ou elle estime nécessaire, soit pour pouvoir avoir une perspective longitudinale sur toute la durée du projet, soit en termes de temps alloué pour procéder à l'évaluation de fin de projet, ou les deux. Dans ce scénario, il n'est pas possible d'effectuer une étude de référence à l'aide d'une méthodologie comparable à celle de l'évaluation finale initialement planifiée. Le temps disponible pour préparer les consultations des intervenants et intervenantes, les visites sur place et les travaux sur le terrain, ainsi que l'analyse des données, pourrait également devoir être considérablement réduit pour respecter la date limite de présentation du rapport. Ces contraintes de temps sont particulièrement problématiques pour un évaluateur ou une évaluatrice qui n'est pas familier-ère avec la région, ni même avec le pays, et qui n'a pas le temps de se familiariser avec les communautés et les organismes qui participent à l'étude et d'instaurer la confiance avec eux. La combinaison des contraintes de temps et des contraintes budgétaires signifie souvent que les évaluateurs et les évaluatrices étrangers-ères ne peuvent séjourner dans le pays que pendant une courte période, ce qui les oblige souvent à recourir à des raccourcis qu'ils et elles reconnaissent comme étant méthodologiquement discutables.

Contraintes budgétaires

Souvent, les fonds pour l'évaluation n'étaient pas inclus dans le budget initial du projet et l'évaluation doit être effectuée avec un budget beaucoup plus faible que celui normalement alloué à ce type d'étude. Par conséquent, il n'est peut-être pas possible d'appliquer les instruments de

collecte de données souhaitables (par exemple, des études de trajectoire ou des enquêtes par sondage), ni d'appliquer les méthodes de reconstitution des données issues de l'étude de référence ou de création de groupes témoins. Le manque de fonds peut également être à l'origine de plusieurs des contraintes de temps évoquées plus haut.

Contraintes liées aux données

Lorsque l'évaluation ne débute qu'à la fin du cycle du projet, il existe généralement peu ou pas de données de référence comparables sur les conditions du groupe cible avant le début du projet. Même si les registres du projet (*project records*) sont disponibles, ils ne sont souvent pas organisés sous la forme requise pour être comparés avant et après l'analyse. Les registres de projet et d'autres données secondaires souffrent souvent de biais systématiques en matière de déclaration ou de mauvaises normes de tenue des dossiers. Même lorsque des données secondaires sont disponibles pour une période proche de la date de début du projet, elles ne correspondent généralement pas entièrement aux populations du projet. Par exemple, les données sur l'emploi ne couvrent que les grandes entreprises, alors que de nombreuses familles investies dans le projet travaillent dans des petites entreprises du secteur informel; les registres scolaires peuvent ne couvrir que les écoles publiques; un autre problème est que les données d'enquête sont souvent agrégées au niveau des ménages, de sorte que l'on ne dispose pas d'informations sur les membres du ménage. Il s'agit là d'un problème particulier pour l'analyse sexo-différenciée.

La plupart des organismes ne s'intéressent qu'à la collecte de données sur les groupes avec lesquels ils travaillent. Ils peuvent également craindre que la collecte d'informations sur les non-bénéficiaires ne crée des attentes en matière de compensation financière ou autre pour ces groupes, ce qui décourage davantage la collecte de données sur un

groupe témoin. Il est également souvent difficile d'identifier un groupe témoin, même si des fonds sont disponibles. De nombreuses zones concernées par le projet évalué présentent des caractéristiques uniques qui font qu'il est difficile de trouver des zones de contrôle comparables. Par exemple, le projet peut concerner toutes les communautés les plus pauvres, ou il a sélectionné toutes les communautés les plus dynamiques, ou n'est organisé que dans des districts où l'appui politique est solide et où l'administration locale s'est engagée à verser des fonds.

Dans d'autres cas, l'impact du projet concerne des sujets sensibles tels que l'autonomisation des femmes, l'utilisation de contraceptifs, la violence familiale ou communautaire ou la corruption, où l'information est difficile à recueillir même lorsque des fonds sont disponibles. Des problèmes de données similaires peuvent survenir lorsque le projet travaille avec des groupes difficiles à atteindre tels que les toxicomanes, les criminels, les minorités ethniques, les migrants et les migrantes, les résidents illégaux et les résident-e-s en situation irrégulière ou, dans certains cas, les femmes.

[…]

2. Le bon, la bête et l'évaluateur : 25 ans d'éthique dans l'American Journal of Evaluation

MICHAEL MORRIS

[Traduit de : Morris, Michael. 2010. « The Good, the Bad, and the Evaluator: 25 Years of AJE Ethics ». *American Journal of Evaluation*, 32(1) : 134-151 (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Au cours du dernier quart de siècle, d'importants progrès ont été accomplis dans le domaine de l'éthique de l'évaluation. Il s'agit notamment de la publication des versions originales et révisées des « Principes directeurs » à l'intention des évaluateurs et des évaluatrices (American Evaluation Association, 1995, 2004), des deuxième et troisième éditions des Normes d'évaluation des programmes (Joint Committee on Standards for Educational Evaluation 1994; Yarbrough *et al.*, 2011), le premier (et jusqu'à présent le seul) texte consacré à l'éthique de l'évaluation des programmes (Newman et Brown, 1996), de chapitres sur l'éthique dans les principaux manuels dans le champ de l'évaluation (par exemple Morris, 2003; Simons, 2006; Wolf, Turne, et Toms, 2009), ainsi que le lancement de la section Défis éthiques de l'*American Journal of Evaluation* (Morris, 1998) et l'ajout d'un module de formation à l'éthique sur le site Web de l'*American Evaluation Association*. Depuis 2000, chaque numéro de l'*American Journal of Evaluation* contient, sous une forme abrégée, les « Principes directeurs » pour les évaluateurs et les évaluatrices. Cet essai explore, d'un point de vue thématique, les contributions des auteurs de l'AJE à notre compréhension de l'éthique de l'évaluation. Cette revue porte

sur tous les numéros *d'Evaluation Practice* (EP) *de* 1986 à 1997, et de l'A*merican Journal of Evaluation* (AJE) de 1998 à juin 2010, avec d'autres sources citées le cas échéant. C'est un voyage intéressant.

S'orienter

Le terme « éthique » a au moins trois significations pertinentes pour cet article (Newman et Brown, 1996 : 20). Au niveau le plus général, l'éthique se réfère aux principes de base du comportement moral – « ce qui est bon et ce qui est mauvais » – qui sont considérés comme s'appliquant largement. Un deuxième sens porte sur les « principes de conduite élaborés par et pour les membres d'une profession particulière » (Morris, 2008 : 2). Enfin, l'étude des valeurs, des croyances et des actions des individus en rapport avec les préoccupations morales représente un troisième sens de l'éthique.

L'étude de l'EP/AJE fournit des exemples des trois utilisations. De temps en temps, un auteur ou une autrice présente un commentaire ou un paragraphe sur l'éthique dans un article qui, autrement, ne vise pas explicitement ce domaine. Par exemple, dans ses recommandations pour les relations avec les médias lors de la diffusion des résultats, Mangano (1991) insiste sur le fait de « ne jamais mentir à un journaliste » (p.119). Hurwith et Sweeney (1995), en discutant de l'utilisation d'images visuelles dans l'évaluation, demandent « Quelles questions éthiques faut-il prendre en considération en ce qui concerne les photographies? » (p.163). Et Morabito (2002), en proposant le rôle de « conseiller organisationnel » comme moyen d'élargir l'influence du processus d'évaluation, note que « l'évaluateur ne devrait pas « conseiller » des individus au sein de l'organisation; c'est le travail de professionnels certifiés…. Le fait d'assumer le rôle de conseiller organisationnel en soi crée des problèmes d'éthique et de frontières » (p.329). De brèves références comme celles-ci avertissent le lecteur ou la lectrice que « quelque chose se trame »

dans le domaine du bien et du mal, mais qu'il n'y aura pas d'exploration approfondie, souvent parce que la question morale en cause est considérée comme évidente (par exemple, mentir est mal), ou que la préoccupation éthique est perçue comme tangente à l'essentiel de l'article.

Il n'est pas surprenant que les questions d'éthique apparaissent le plus souvent dans EP/AJE lorsque les auteurs et autrices utilisent des principes et des normes professionnels particuliers dans leur analyse ou soulèvent des questions qui ont des implications majeures pour ces principes et normes (deuxième sens de l'éthique). Dans la première catégorie rentrent la plupart des essais de la série *Défis éthiques*, dans lesquels les commentateurs et commentatrices sont prié-e-s d'examiner un scénario d'évaluation du point de vue des principes directeurs à l'intention des évaluateurs et des évaluatrices et/ou des standards de l'évaluation de programme. Les articles qui examinent l'évaluation dans des environnements hautement politisés comme le gouvernement fédéral (Chelimsky, 2008), soutiennent que les évaluateurs et les évaluatrices ont la responsabilité de servir de critiques moraux des programmes sociaux (Schwandt, 1992) ou de démontrer que l'évaluation devrait faire progresser la cause de la justice sociale en incluant de façon proactive les intervenants et intervenantes marginalisé-e-s (Mertens, 1999, 2001).

Enfin, un petit nombre d'enquêtes empiriques menées dans PE/AJE ont porté sur des questions importantes en matière d'éthique de l'évaluation. Walker, Hoggart et Hamilton (2008) se sont servis d'observations, d'entretiens en face-à-face et téléphoniques et de groupes de discussion pour explorer les perceptions que les usagers et usagères ainsi que le personnel avaient de la mise en œuvre du consentement éclairé dans le cadre de l'expérimentation d'une politique sociale à grande échelle au Royaume-Uni. Un autre exemple est Penuel, Sussex, Korbak et Hoadley (2006), qui ont interviewé des directeurs/-trices d'école, des enseignant-e-s et des chef-fe-s d'établissement pour recueillir leurs réactions à l'utilisation potentielle de l'analyse des réseaux sociaux comme méthode d'évaluation des données dans les écoles. La protection de la vie privée

est apparue comme une préoccupation importante, ce qui, à son tour, pose un défi éthique aux évaluateurs et évaluatrices qui souhaitent utiliser cette méthodologie dans leur travail.

La rareté des recherches sur l'éthique signalées dans l'EP/AJE au fil des ans ne signifie pas nécessairement que ce sujet soit négligé. En général, on manque dans la littérature évaluative d'études empiriques du champ, bien que la situation s'améliore. Comme l'ont fait remarquer Henry et Mark (2003), « il y a une grave pénurie de preuves rigoureuses et systématiques qui peuvent guider l'évaluation ou que les évaluateurs et les évaluatrices puissent utiliser pour renforcer leur réflexivité ou améliorer leur prochaine évaluation » (p.63). À cet égard, du moins, l'éthique de l'évaluation voyage en bonne compagnie.

Bien entendu, de nombreux articles de l'EP/AJE ne traitent pas explicitement de l'éthique en fonction de ses trois significations. À part Penuel *et al.* (2006), cela est particulièrement vrai pour les articles mettant l'accent sur la méthodologie et les techniques d'analyse des données. Seule la crainte de relancer les guerres qualitatives-quantitatives m'empêchent de proposer la *Loi des développements incompatibles* de Morris : « Les coefficients de régression et les analyses éthiques fleurissent rarement dans le même jardin. » Cela dit, la pépinière EP/AJE a cultivé assez de contenu éthique pour justifier un passage en revue des grands thèmes abordés. Cette analyse sera organisée en fonction des cinq domaines définis dans les « Principes directeurs » de l'AEA pour les évaluateurs et évaluatrices : investigation systématique, compétence, intégrité/honnêteté, respect des personnes et responsabilités en matière de bien-être général et public. Bien que ces domaines se chevauchent de multiples façons (par exemple, les évaluateurs/-trices compétent-e-s sont plus susceptibles de mener un examen systématique valide que les évaluateurs/-trices incompétent-e-s), ils sont suffisamment distincts pour en faire un cadre utile à des fins d'analyse. L'examen se terminera par quelques conclusions et

recommandations modestes qui, dans l'esprit des bonnes pratiques d'évaluation, sont destinées à s'inspirer, dans une large mesure, des données qui ont été examinées.

Investigation systématique

Le premier principe directeur stipule que « les évaluateurs et les évaluatrices mènent des investigations systématiques fondées sur des données ». Au cœur de ce principe se trouve l'idée que « les évaluateurs et les évaluatrices ont la responsabilité de concevoir et de mener des études qui sont techniquement solides et qui répondent adéquatement aux questions d'évaluation qui sous-tendent le projet » (Morris 2008 : 8-9). Parmi les cinq « Principes directeurs », celui-ci porte le plus directement sur les décisions méthodologiques prises lors de l'évaluation, bien que ce principe ne rende aucun jugement particulier favorisant certaines méthodes par rapport à d'autres.

L'importance éthique de la responsabilité méthodologique est clairement illustrée dans un premier article de Lipsey (1988). Dans sa critique de la rigueur des recherches expérimentales et quasi expérimentales, très quantitatives dans leur orientation, il a conclu que « nous ne pratiquons pas l'évaluation, nous la pratiquons mal » (p.22) et que « ce que nous faisons actuellement…. est souvent pire que simplement inutile, c'est activement nuisible » (p.6). Son essai a suscité une vive réponse de la part de Hennessy et Sullivan (1989), qui se sont penchés sur les contraintes pratiques auxquelles les évaluateurs et les évaluatrices doivent s'adapter lorsqu'ils et elles effectuent des recherches dans des environnements organisationnels réels qui englobent différents styles décisionnels. Au fur et à mesure que le débat se poursuivait, il est devenu évident que, dans une large mesure, les auteurs/-ices se parlaient sans se comprendre (comme c'est souvent le cas dans de tels litiges). Le principal objectif de Lipsey était d'attirer l'attention sur l'application erronée de méthodes prestigieuses, tandis que Hennessy et Sullivan s'intéressaient principalement à expliquer les circonstances qui pouvaient conduire à

une telle application (Hennessy et Sullivan, 1990; Lipsey, 1990). Finalement, Hennessy et Sullivan semblent plaider « non coupable », au nom de la plupart des évaluateurs et évaluatrices, à l'accusation de faute professionnelle. Mais le message de Lipsey reste un message puissant; à savoir que les évaluateurs et les évaluatrices ne peuvent échapper à l'examen éthique en se cachant derrière des méthodologies de haut niveau, une vérité qui nous a été rappelée par les controverses récentes entourant les commandes d'évaluations randomisées.

On peut également trouver des cas où les auteurs de l'EP/AJE portent leur attention sur la pertinence méthodologique d'évaluations spécifiques (ou de composantes d'évaluation); [...] Dans l'ensemble, ces analyses soulignent que les méthodologies ne se choisissent pas elles-mêmes; les évaluateurs et les évaluatrices ainsi que les autres intervenantes et intervenants prennent des décisions qui ont des conséquences sur la qualité et l'utilité de la recherche menée et dont il est rendu compte. Dans certains cas, les auteurs et les autrices critiquent fortement ces choix. Dans leur examen de 12 grandes évaluations fédérales de programmes au sein du ministère de la Santé et des Services sociaux des États-Unis, par exemple, Howell et Yemane (2006) affirment que « le plus grand défaut des évaluations examinées a été l'absence d'une étude de l'impact et des problèmes liés au choix de groupes de comparaison appropriés pour celles qui utilisaient des approches quasi-expérimentales » (p.232), et ils concluent que « malgré les investissements importants dans ces programmes, les évaluations n'ont à ce jour donné aucune preuve solide quant à savoir si ces programmes valaient ce qu'ils avaient coûté en financements fédéraux » (p.234). Dans un autre examen au niveau fédéral, Renger (2006) a abordé l'utilisation des modèles logiques par le Bureau des professions de la santé. Il a constaté que « l'agence n'a pas utilisé au mieux le processus de modélisation logique et n'a pas observé les meilleures pratiques dans la création de ces modèles » (p.461) et se demande si « peut-être le monde [de l'évaluation] devrait exercer un plus grand contrôle sur l'application et la mise en œuvre de ses méthodes afin

d'améliorer la probabilité de produire des évaluations de la qualité. *Le fait de ne pas le faire pourrait saper la crédibilité et l'intégrité du champ de l'évaluation* » (p.462, c'est moi qui souligne).

Contrairement aux articles relatifs à la bonne application des méthodes conventionnelles, certain-e-s soulèvent des questions quant à la capacité des approches évaluatives de rendre justice au principe d'examen systématique. L'une des cibles les plus importantes dans ce contexte a été l'évaluation émancipatrice (*empowering evaluation*), qui met l'accent sur « (1) la fourniture aux parties prenantes du programme d'outils d'évaluation de la planification, de la mise en œuvre et de l'auto-évaluation de leur programme, et (2) l'évaluation de l'intégration dans le cadre de la planification et de la gestion du programme ou de l'organisation » (Wandersman, Snell-Johns *et al.*, 2005 : 28). Depuis la formulation initiale par Fetterman de l'évaluation émancipatrice dans son discours présidentiel de 1993 de l'AEA (Fetterman, 1994), l'approche a été critiquée depuis de multiples perspectives. Stufflebeam (1994) a affirmé que « l'évaluation émancipatrice ignore essentiellement les *Standards* [de l'évaluation de programme] et…. va dans le sens de faire de l'évaluation un exercice massif de relations publiques et de processus de groupe » (p.333, italiques présentes dans l'original; voir aussi Lackey, Moberg, et Balistrieri, 1997). Au fil des ans, les critiques de Scriven ont probablement été les plus énergiques et les plus persistantes et sont directement liées au principe de l'examen systématique. Il affirme que l'évaluation émancipatrice « manque de crédibilité parce qu'il s'agit d'une auto-évaluation, et tout le monde sait que l'auto-évaluation est sujette au biais majeur de la surévaluation de soi-même et de son travail » (2005 : 415). Il conclut que l'approche « ne répond pas aux principales normes minimales…. de la validité, de la crédibilité et de l'éthique » et qu'il constitue « une auto-évaluation amatrice », ce qui représente « un modèle absurde pour une évaluation sérieuse » (p.431; voir aussi Scriven, 1997). En réponse, Fetterman et Wandersman (2007) soutiennent qu'ils ont constaté que « de nombreuses évaluations émancipatrices étaient hautement critiques à l'égard de leurs propres opérations [de programme] » (p.184) et

Wandersman et Snell-Johns (2005) soulignent qu'« un rôle crucial d'un évaluateur émancipateur consiste à s'assurer que le système d'auto-évaluation recueille des renseignements exacts » (p.424).

[…]

Des études expérimentales sur la prise de décision des évaluateurs et des évaluatrices peuvent également éclairer les dimensions éthiques de l'examen systématique. Dans un excellent exemple de cette approche, Azzam (2010) a désigné au hasard des évaluateurs et des évaluatrices pour qu'ils et elles reçoivent une réponse acceptant ou rejetant l'approche qu'elles et ils proposaient pour l'étude d'un programme scolaire fictif. Les commentaires ont été formulés par trois parties prenantes différentes : le chef d'établissement (le décideur du programme), les tuteurs-enseignants et les tutrices-enseignantes (les exécutants du programme) et les parents (les bénéficiaires du programme). On a ensuite demandé aux évaluateurs et aux évaluatrices s'ils et elles souhaitaient modifier leur proposition. Les résultats indiquent que les évaluateurs et les évaluatrices étaient plus susceptibles de la modifier lorsque des objections étaient soulevées par des parties prenantes ayant plus de pouvoir ou d'influence sur les facteurs logistiques ayant une incidence sur l'évaluation (c'est-à-dire les chefs d'établissement). Certes, on peut changer une approche ou un modèle suite à un réexamen de bonne foi de la qualité de l'approche ou du modèle original. Toutefois, il peut aussi s'agir d'une capitulation par laquelle un modèle plus faible est choisi en réponse à la pression perçue de parties prenantes puissantes. Cette dernière circonstance pourrait menacer la capacité de l'évaluation à atteindre un niveau d'adéquation technique conforme au principe de l'examen systématique. En effet, dans son analyse des raisons professionnelles justifiant le refus d'un contrat d'évaluation, Smith (1998) recommande qu'un contrat soit refusé si le travail souhaité n'est pas possible à un niveau de qualité acceptable (p.179). Il serait difficile de trouver un énoncé plus succinct de la responsabilité éthique de l'évaluateur à l'égard de la méthodologie. Un jour, lorsque l'utopie évaluatrice adviendra, les termes « qualité » et « niveau acceptable » seront définis de la même manière par toutes les parties

prenantes. D'ici là, les évaluateurs et les évaluatrices doivent continuer de faire les vérifications nécessaires dans la négociation d'un terrain méthodologique contesté avec les client-e-s, les autres parties prenantes et leurs collègues évaluateurs/-trices.

Compétence

Selon le principe de compétence, « les évaluateurs et les évaluatrices fournissent une performance compétente aux parties prenantes ». Cette déclaration préliminaire minimaliste est suivie de détails portant sur des questions d'éducation, d'expérience, de compétence culturelle, d'expertise pertinente et de perfectionnement professionnel. [...]

Vues ensemble, que nous disent ces analyses?

- L'application de lignes directrices professionnelles générales à des problèmes d'évaluation particuliers n'est pas un processus mécanique qui produit des résultats uniformes. Lorsqu'ils et elles sont libres, les commentateurs et les commentatrices peuvent varier considérablement dans les principes et les normes qu'ils et elles citent comme étant les plus pertinents pour un cas donné.
- Compte tenu de l'observation précédente, il ne faut pas s'étonner que les auteurs/-trices puissent différer, parfois fortement, dans leurs recommandations quant à la façon dont un évaluateur ou une évaluatrice devrait relever un défi éthique particulier.
- Les commentateurs et les commentatrices sont sensibles aux exigences contradictoires que les lignes directrices professionnelles peuvent présenter à l'évaluateur ou à l'évaluatrice et peuvent différer quant à la façon dont ils et elles règlent ces conflits.
- Les commentateurs et les commentatrices vont parfois au-delà des « Principes directeurs » et des Standards d'évaluation de programme pour chercher des conseils et des justifications pour leurs réponses

aux conflits éthiques. Des pratiques commerciales normalisées, des accords négociés, des contrats officiels et des cadres moraux généraux comme une *éthique du care* ont été invoqués (Smith, 2002).

Le message général qui se dégage ici est que les principes et les standards professionnels constituent un cadre précieux pour identifier et conceptualiser les questions d'éthique dans l'évaluation, mais qu'ils ne se traduisent pas facilement en un manuel pratique pour prendre des décisions et des mesures précises. Ce n'est pas une idée nouvelle; House (1995) et Rossi (1995) ont présenté des observations similaires lorsque la version originale des « Principes directeurs » a été publiée, et Mabry (1999) a exploré cette question en profondeur dans son article influent de l'AJE « Éthique circonstancielle », dans lequel elle a observé que « les couches de complexité et de nuance dans le comportement humain et l'organisation sociale submergent les déclarations formelles [de normes et de principes] avec des situations imprévues » (p.201).[…]

Bien entendu, la compétence de l'évaluateur ou de l'évaluatrice ne se limite pas au simple fait de relever les défis éthiques spécifiques que l'on rencontre dans son travail. L'identification des multiples éléments de connaissances et de savoir-faire qui sous-tendent cette compétence est une tâche qui a fait l'objet d'une attention beaucoup plus soutenue ces dernières années, ce qui a conduit à l'élaboration d'une taxonomie détaillée et empiriquement formulée en 2005 (Stevahn *et al.*, 2005). Les principales catégories de cette taxonomie sont la pratique professionnelle (qui comprend les normes et l'éthique), l'examen systématique, l'analyse de la situation, la gestion de projet, la pratique réflexive et la compétence interpersonnelle. On peut soutenir qu'une telle taxonomie est inhérente à une « pensée » éthique, c'est-à-dire que les personnes qui se disent évaluateurs ou évaluatrices devraient posséder des niveaux de compétence minimaux dans tous les domaines précis identifiés ou, du moins aux fins d'un projet donné, s'associer à d'autres évaluateurs ou évaluatrices qui possèdent les compétences dont ils ne disposent pas. Au niveau formel, toutefois, l'AEA n'a pas officiellement approuvé cette

taxonomie ou aucune autre taxonomie, et n'a pas non plus adopté de procédure de certification ou de certification des évaluateurs et des évaluatrices, bien que cette dernière question ait fait l'objet d'un débat approfondi (Altschuld, 1999a, 1999b; Bickman, 1999; Smith, 1999; Worthen, 1999). Ainsi, *Caveat Emptor (que l'acheteur soit vigilant*) demeure le meilleur conseil que le domaine puisse offrir aux parties prenantes à la recherche d'évaluateurs et d'évaluatrices compétent-e-s, éthiques ou autres. Dans ce contexte, il n'est pas choquant que Dewey, Montrosse, Schroëter, Sullins et Mattox (2008) aient constaté que certaines compétences souhaitées par les employeurs et employeuses d'évaluateurs et d'évaluatrices (par exemple, les compétences en gestion interpersonnelle et de projet) ne correspondent pas aux compétences que les demandeurs et demandeuses d'emploi acquièrent en troisième cycle ou que dans l'étude de Taut et Alkin (2003) sur les perceptions du personnel de programme quant aux obstacles à la mise en œuvre de l'évaluation, la compétence sociale de l'évaluateur ou de l'évaluatrice soit apparue comme un enjeu essentiel.

Le domaine spécifique lié aux compétences dont l'importance a probablement le plus augmenté au cours des deux dernières décennies et demie est celui de la compétence culturelle. Les « Principes directeurs » initiaux ne mentionnent même pas la culture dans la définition du principe de compétence. Dans la version révisée, la compétence culturelle figure en bonne place. De même, dans l'édition de 1994 des Standards d'évaluation de programme, on cherche en vain dans le glossaire une mention sur la « culture ». Dans la troisième édition, les définitions de « compétence culturelle » et d'« évaluation attentive à la culture » (*culturally responsive evaluation*) sont parmi les plus détaillées dans l'ouvrage (Yarbrough *et al.*, 2011 : 286). De plus, les auteurs et les autrices notent qu'un changement majeur dans la troisième édition est le traitement explicite des [thèmes] de la culture et du contexte et observent que ces facteurs « ont de profondes influences sur la façon dont les évaluateurs se préoccupent de et renforcent l'utilité, la pertinence, l'exactitude et la redevabilité » (p.xiii-xiv). Avec la publication

en 2009 de la revue de littérature sur l'évaluation interculturelle de Chouinard et Cousins, nous pouvons conclure en toute sécurité que la compétence culturelle est désormais un membre en bonne et due forme –et influent qui plus est– de l'actuel *zeitgeist* de l'évaluation. […]

EP/AJE a publié sa part d'articles sur les facteurs culturels dans l'évaluation, dont l'examen détaillé dépasse le cadre de cet essai (Birman, 2007; Botcheva, Shih, et Huffman, 2009; Botcheva, White, et Huffman, 2002; Clayson *et al.*, 2002; Cooksy, 2007; Fitzpatrick, 2007; Hopson ,1999, 2001; House, 1999; Jaycox *et al.*, 2006; Kirkhart, 1995; Mertens, 2007; Patton, 1999; Slaughter, 1991; Stanfield, 1999; Williams, 1991). Pour nos besoins, le point crucial est que la réceptivité culturelle est considérée comme un impératif éthique pour les évaluateurs et les évaluatrices. Cette réactivité est importante non seulement parce qu'elle peut améliorer la qualité de la recherche menée (examen systématique), mais aussi parce qu'elle honore la dignité des intervenants et intervenantes avec lesquels l'évaluateur ou l'évaluatrice interagit (Respect pour les personnes) et qu'elle est pertinente pour l'intérêt public (Responsabilités pour le bien-être général et le bien-être public). Toutefois, la littérature pertinente indique également que la compétence culturelle n'est pas facile à atteindre et qu'elle exige un niveau de réflexivité qui peut être difficile à maîtriser sans orientation structurée (Jewiss et Clark-Keefe, 2007). En d'autres termes, le travail des évaluateurs et des évaluatrices qui prennent au sérieux les compétences culturelles ne peut être « simplement expliqué ».

Intégrité/honnêteté

Le principe directeur de l'intégrité et de l'honnêteté stipule que « les évaluateurs et les évaluatrices font preuve d'honnêteté et d'intégrité dans leur propre comportement et s'efforcent d'assurer l'honnêteté et l'intégrité de l'ensemble du processus d'évaluation ». Les évaluatrices et

les évaluateurs doivent être proactifs-ves dans la création d'un environnement qui permet aux parties prenantes une compréhension complète de la planification, des opérations et des résultats de l'évaluation. Les conflits d'intérêts, les changements apportés à la conception de l'évaluation au cours de la mise en œuvre et le fait de rendre fidèlement compte des constats effectués figurent parmi les questions abordées dans le principe de l'intégrité et de l'honnêteté.

Compte tenu de l'importance de ces questions, il est curieux qu'une plus grande attention n'ait pas été accordée à ces questions dans la littérature sur l'évaluation, y compris l'EP/AJE. Bien entendu, les évaluateurs et les évaluatrices peuvent avoir l'impression que, dans l'ensemble, elles et ils s'acquittent plutôt bien des tâches liées à ce principe, évitant ainsi la nécessité d'un grand nombre de discussions explicites. Les recherches indiquent toutefois que l'un des problèmes éthiques les plus fréquents rencontrés par les évaluatrices et les évaluateurs est d'être soumis à des pressions de la part d'une partie prenante afin de dénaturer les résultats d'une étude (Morris et Clark, 2009; Morris et Cohn, 1993; Turner, 2003). Dans ce contexte, Chelimsky a vigoureusement, éloquemment et constamment souligné combien il est important que les évaluatrices et les évaluateurs ne se plient pas aux pressions politiques qui peuvent saper leur crédibilité et leur indépendance au cours des phases de conception, de mise en œuvre, de l'établissement des constats et de diffusion de l'étude (par exemple Chelimsky, 1995, 2008; Chelimsky, Cordray, et Datta, 1989; voir aussi Hedrick, 1988). Elle fait remarquer que « toutes les choses ne peuvent pas être équilibrées et sujettes à compromis dans l'évaluation. Les constats sont des constats, et le fait qu'ils soient soutenus par les données signifie qu'ils sont soutenus par les données. L'intégrité de notre travail doit toujours être défendable – et défendue avec force – dans un environnement toujours politique » (2008 : 412). Chelimsky insiste sur la nécessité de faire preuve de courage face à ces défis : « il faut du courage pour refuser aux commanditaires les réponses qu'ils ou elles veulent entendre, pour résister à l'idée de devenir « un-e membre de l'équipe », pour contester les mythes avec lesquels tout le monde est à l'aise, pour

lutter contre les intrusions inappropriées dans le processus d'évaluation » (1995 : 220). L'implication est claire : parfois, la seule raison de faire une bonne chose c'est parce que c'est la bonne chose à faire. C'est dangereux, et de puissantes rationalisations sont souvent facilement disponibles pour justifier la poursuite d'une voie de moindre résistance… et de risque. Les évaluateurs et évaluatrices internes, qui sont intégré-e-s dans des réseaux de dynamiques politiques et hiérarchiques transversales au sein de l'organisation, sont particulièrement vulnérables à ces forces et à ces rationalisations. Heureusement, Chelimsky et d'autres ont offert pléthore de conseils sur la façon de conceptualiser et de gérer le processus d'évaluation d'une manière qui maximise la probabilité que des études honnêtes et crédibles puissent être produites sans avoir à compter sur l'héroïsme personnel (Chelimsky, 2008; Datta, 2000; Iriti, Bickel et Nelson, 2005).

Le mésusage des évaluations est l'autre grand sujet relevant du principe de l'intégrité et de l'honnêteté, qui n'a pas fait l'objet d'une attention particulière dans l'EP/AJE. Le principe stipule que « dans des limites raisonnables, ils [les évaluateurs et les évaluatrices] devraient tenter de prévenir ou de corriger l'utilisation abusive de leur travail par d'autres ». Il est certain que l'utilisation abusive a souvent lieu à un moment où la capacité de l'évaluateur ou de l'évaluatrice d'influencer les événements est au mieux modeste. Des enquêtes indiquent qu'environ un quart des évaluatrices et des évaluateurs ont connu des problèmes majeurs d'utilisation abusive intentionnelle (Fleischer et Christie, 2009; Preskill et Caracelli, 1997; Torres, Preskill, et Piontek, 1997), proportion qui semble être restée relativement stable au fil du temps. Comme c'est le cas pour les pressions exercées pour déformer les conclusions, il est préférable d'empêcher la tentative d'utilisation abusive plutôt que d'avoir à faire face au problème une fois qu'il s'est produit (voir Stevens et Dial, 1994). Néanmoins, il serait utile d'étudier plus systématiquement la façon dont les évaluateurs et les évaluatrices réagissent aux épisodes d'utilisation abusive lorsqu'ils se produisent. Par exemple, à quelle fréquence les évaluateurs et les évaluatrices prennent-ils et elles des mesures pour

s'attaquer réellement aux mauvais usages? Quels sont les facteurs qui influent sur leur décision d'intervenir ou non? Lorsque les évaluatrices et les évaluateurs réagissent à l'utilisation abusive, quelles sont les formes que prennent leurs actions? Des typologies d'utilisation, de non-utilisation, d'utilisation abusive et d'utilisation légitime sont disponibles pour aider à encadrer ces recherches (Cousins, 2004; Patton, 1988). Pour sûr, il y a bien quelques thèses de doctorat qui attendent d'être menées à ce sujet.

Le respect des personnes

Selon le principe directeur du respect des personnes, « les évaluatrices et les évaluateurs respectent la sécurité, la dignité et l'estime de soi des répondant-e-s, des participant-e-s au programme, des client-e-s et des autres parties prenantes de l'évaluation ». Bien que ce principe englobe l'éventail des sujets traditionnellement associés au traitement éthique des sujets humains (par exemple le consentement éclairé, la confidentialité/anonymat, les incitations à la participation, l'analyse des risques et avantages et la justification de la randomisation), le terrain éthique est compliqué par la multitude de groupes de parties prenantes qui peuvent être pertinents pour une évaluation donnée. Faire preuve de respect envers toutes les personnes dans ces circonstances peut être une tâche redoutable. L'évaluation a énormément bénéficié du travail effectué dans d'autres domaines (par exemple, la médecine et la psychologie) sur ces questions. Un défi constant consiste à adapter les connaissances générées par cette recherche au terrain distinctif de l'évaluation de programme, tout en s'appuyant sur celles-ci (Barkdoll, 1992).

Diverses contributions à EP/AJE examinent des questions spécifiques dans le domaine du respect pour les personnes. Les analyses qui, explicitement ou implicitement, abordent certains aspects de la notion de « ne pas nuire » dans des contextes d'évaluation complexe sont

particulièrement intéressantes. Par exemple, que se passe-t-il lorsque les évaluateurs ou les évaluatrices définissent de façon étroite le fait d'être sans-abri, ce qui donne lieu à des constats qui concordent avec les perceptions stéréotypées de cette population? Comme l'affirment Johnson, Mitra, Newman et Horm (1993), une conséquence est que les politiques sociales fondées sur cette définition pourraient causer plus de dommages que de bien, les besoins de certains sous-groupes fortement marginalisés n'étant pas pris en compte. English (1997) met l'accent sur les limites du consentement éclairé pour protéger les intérêts des groupes de parties prenantes défavorisées et minoritaires dans l'évaluation des services destinés aux personnes handicapées. Ferris (2000) fait état des difficultés rencontrées par les évaluateurs et par les évaluatrices pour protéger la confidentialité et obtenir le consentement éclairé dans une étude sur les fournisseurs de services d'avortement au Canada, qui peuvent être victimes de harcèlement. Et dans l'étude exemplaire réalisée par Walker *et al.* (2008) sur le programme de maintien et d'avancement dans l'emploi (*Employment Advancement & Retention*, ERA) au Royaume-Uni, ils ont constaté que « le personnel surestimait assez fréquemment la mesure dans laquelle les clients et les clientes comprenaient quels étaient les services et les soutiens fournis au titre de l'ERA, la nature de la recherche et l'importance de l'assignation aléatoire » (p.170), résultat qui était au moins partiellement attribuable à la « pression exercée sur le personnel pour maximiser le nombre de volontaires pour participer à l'expérimentation » (p.165). Il est difficile d'imaginer que des études aussi approfondies sur le consentement éclairé dans d'autres évaluations ne trouveraient pas de preuves des types de dynamiques identifiées par Walker et ses collègues.

L'un des principaux thèmes de la plupart de ces études est la vulnérabilité des parties prenantes qui ont peu de pouvoir dans les évaluations, ainsi que les mesures proactives, souvent créatives, nécessaires pour prévenir ou minimiser les dommages et promouvoir une implication qui ait du sens (voir aussi Alexander et Richman, 2008; Balch et Mertens, 1999; Campbell, Adams, et Patterson, 2008; Gill, 1999; Mertens, 1999). En effet,

la sensibilité accrue du champ de l'évaluation aux aspects de la pratique professionnelle dans ce domaine au cours des deux dernières décennies a été remarquable.

Les répercussions éthiques des méthodes nouvelles ou nouvellement appliquées sont également examinées, et une attention soutenue continue d'être accordée à des sujets classiques tels que les comités d'éthique (*Institutional Review Boards*) (Cooksy, 2005) et le consentement actif par opposition au consentement passif (Johnson *et al.*, 1999; Leakey *et al.*, 2004). Par exemple, les données des entreprises (Spicer *et al.*, 2004), l'évaluation des praticiennes et des praticiens (Shaw et Faulkner, 2006), l'analyse des réseaux sociaux (Penuel *et al.*, 2006) et les systèmes d'information géographique (Renger *et al.*, 2002) contiennent toutes des caractéristiques qui peuvent compromettre la vie privée des parties prenantes si elles sont utilisées sans précautions appropriées.

Un domaine important qui n'a guère été exploré dans EP/AJE est celui de l'analyse risques-avantages, où les risques encourus par les divers intervenants de l'évaluation sont comparés aux avantages qu'ils sont susceptibles de tirer de l'évaluation. Des études de cas d'évaluations controversées (et non controversées) du point de vue du risque et des avantages amélioreraient notre compréhension des compromis parfois complexes dont les évaluateurs et les évaluatrices devraient tenir compte lorsqu'ils et elles conçoivent des études qui respectent le principe du respect des personnes.

Responsabilités en matière d'intérêt général et intérêt public (General and public welfare)

Le cinquième principe directeur indique que « les évaluatrices et les évaluateurs expriment et tiennent compte de la diversité des intérêts et des valeurs généraux et publics qui peuvent être liés à l'évaluation ».

Il s'agit à bien des égards du principe directeur le plus intrigant, étant donné qu'il implique, sans le définir, des concepts idéologiques aussi lourds de sens que « le bien public », « l'intérêt public » et « ce qui est bon pour le public » (*public good*). Voyez, par exemple, les politiques sociales et économiques soutenues par ceux qui se trouvent à divers endroits du continuum politique gauche-droite. Ou, comme le dit Smith (1998), « alors que tous les évaluateurs et toutes les évaluatrices accepteraient de ne pas faire de mal », il y a sans doute moins d'accords sur la façon de « promouvoir le bien public » (p.184). Par conséquent, il n'est pas surprenant que, dans son examen des commentaires sur les défis éthiques, Datta (2002) fasse remarquer que « les désaccords les plus manifestes ont tendance à porter sur le rôle de l'évaluateur ou de l'évaluatrice, particulièrement [dans] des situations où les évaluateurs et les évaluatrices pourraient se considérer comme des justiciers ou des justicières (*social avengers*) » (p.196). Implicitement, le principe des responsabilités encourage les évaluatrices et les évaluateurs à réfléchir à leur vision du bien public et à tenir compte des répercussions de cette vision sur les évaluations qu'elles et ils mènent. […]

Dans l'ensemble, ces constats suggèrent que de nombreux évaluateurs et de nombreuses évaluatrices considèrent avec ambivalence les questions soulevées par le principe de responsabilité, même au niveau le plus général et le plus abstrait, où l'on peut s'attendre à ce que l'approbation soit élevée. Toutefois, l'ambivalence n'a en aucune façon entraîné un manque de contributions à l'EP/AJE directement liées à ce principe. Une grande partie de cette analyse reflète les différences de compréhension à l'égard du concept de plaidoyer. […]

Sur le plan conceptuel, le plaidoyer pour des recommandations issues de l'évaluation n'est pas la même chose que d'agir en tant que défenseur ou défenseuse des programmes. La grande majorité de la communauté de l'évaluation semble rejeter ce dernier rôle. En se fondant sur son examen sélectif de la littérature, Datta (1999) conclut que « des évaluateurs et des évaluatrices différent-e-s conviennent que l'évaluateur/-trice ne devrait pas être un-e défenseur/-euse (ou vraisemblablement un-e adversaire)

d'un programme particulier, c'est-à-dire prendre parti, ou d'une position préconçue de soutien (ou de suppression) » (p.84). Dans l'étude de Fleischer et Christie (2009), 17 % seulement de leur échantillon d'évaluateurs et d'évaluatrices étaient d'avis que « devenir défenseurs ou défenseuses de programmes » était un atout approprié pour les évaluateurs et les évaluatrices. Si des recommandations d'amélioration d'un programme découlent d'une évaluation, à quels égards la défense de ces recommandations diffère-t-elle de la défense de l'existence du programme lui-même? Ou examinez l'analyse d'Eddy et Berry (2009), qui prétend que, dans certains cas, le respect du principe des responsabilités obligerait l'évaluateur ou l'évaluatrice à recommander la suppression d'un programme insatisfaisant, une mesure que l'évaluateur ou l'évaluatrice devrait considérer comme « une occasion d'exercer sa responsabilité professionnelle à l'égard de la profession d'évaluation et du bien public » (p.375). Cela représente-t-il un plaidoyer? Ou ne parlerait-on de plaidoyer que si l'évaluatrice ou l'évaluateur faisait pression pour que le programme soit arrêté d'une manière qui irait au-delà du simple partage de recommandations avec les principales parties prenantes à travers des habituels rapports et réunions?

[…]

[La] perspective défendue par Schwandt (1992, 2008), est que l'évaluatrice ou l'évaluateur a un rôle central à jouer en tant que critique social. Dans la pratique d'évaluation « moralement engagée » (Schwandt, 1992), l'évaluateur ou l'évaluatrice est chargé-e d'examiner, en collaboration avec les intervenant-e-s, « l'intérêt social des divers résultats et effets des politiques et des programmes » (2008 : 146). En d'autres termes, il ne suffit pas d'accomplir simplement la tâche technique de déterminer ce qu'un programme a accompli ou n'a pas réalisé. Il faut également tenir compte de la légitimité sociale et morale fondamentale des objectifs recherchés. Cette exploration repose sur une vision du bien public qui n'est pas élaborée isolément par l'évaluateur ou l'évaluatrice; il est plus probable qu'elle résulte du genre de processus de réflexion que House et Howe (1999) ont proposé pour l'évaluation démocratique délibérative. En

théorie, cette conceptualisation de l'évaluation honore le principe des responsabilités en articulant une notion de société juste suffisamment détaillée pour s'appliquer au programme en question. L'évaluateur ou l'évaluatrice ne se contente plus de demander : « Le programme fonctionne-t-il bien? » Au contraire, « Le programme fait-il X suffisamment bien? » devient une requête légitime et nécessaire.

Il est difficile de déterminer dans quelle mesure cette vision du rôle de l'évaluatrice ou de l'évaluateur a eu une incidence directe sur les pratiques d'évaluation. Compte tenu des travaux précédemment cités sur l'opinion des évaluatrices et des évaluateurs vis-à-vis d'un agenda de justice sociale pour le champ de l'évaluation, on ne s'attendrait pas à ce que cette influence soit nécessairement grande. Toutefois, il existe des récits dans lesquels les évaluatrices et les évaluateurs jouent ce rôle de critiques sociaux (parfois frustré-e-s). Greene et Lee (2006), par exemple, décrivent les difficultés qu'il et elle ont rencontrées dans l'évaluation d'une initiative globale de réforme scolaire dans une école primaire, où il et elle ont conclu que la réforme elle-même était mal conçue et qu'elle aurait probablement dû aborder d'autres questions. L'auteur et l'autrice ont eu du mal à faire face à cette réalité, à savoir s'il et elle devaient tenter de recentrer l'évaluation sur ces autres questions et par qui une telle réorientation serait légitimée. Le fait de proposer une nouvelle orientation pour l'évaluation représenterait, dans un sens important, un changement de rôle pour les évaluateurs. Et comme le font remarquer Greene et Lee, le changement de rôle dans l'évaluation n'est pas quelque chose que vous faites sans permission. Sur le plan logistique et éthique, vous ne pouvez pas simplement vous déclarer critique social et passer ensuite à l'étape 2 de l'évaluation.

Ryan (2002) puis Lee et Walsh (2004) se sont également montrés sensibles à cette question dans leurs critiques, respectivement, sur les systèmes de responsabilisation en matière d'éducation et sur les définitions de la qualité dans les programmes visant la petite enfance. Dans les deux cas, elles et ils examinent les conséquences sociales de l'utilisation des critères de performance prédominants dans l'évaluation. Et dans les deux

cas, elles et ils demandent l'inclusion de diverses parties prenantes dans les délibérations portant sur ces critères. De même, Friedman, Rothman et Withers (2006) encouragent l'utilisation de « Why Dialogues » pour examiner les valeurs qui sous-tendent les objectifs des parties prenantes en matière de programmes. L'accent n'est pas principalement mis sur ce que sont les objectifs, mais sur les raisons qui les sous-tendent. Plus l'enquête sera approfondie, plus il y aura de chances que les valeurs fondamentales pertinentes pour le bien public seront mises en évidence et que l'évaluateur ou l'évaluatrice pourra s'appuyer dessus de façon productive.

L'inconfort avec un modèle d'évaluation moralement engagée vers la critique sociale est peut-être résumé de façon très succincte dans l'affirmation de Shadish (1994) selon laquelle « la justice sociale est mieux servie.... en permettant aux parties prenantes elles-mêmes d'orienter la sélection des critères de mérite dans l'évaluation » (p.357). Une telle position de « courtier/-ère honnête » ne laisse pas beaucoup de place (le cas échéant) à l'évaluateur ou à l'évaluatrice pour défendre sa propre vision du bien public. Il est intéressant de se demander si des données pourraient résoudre la question soulevée par l'énoncé de Shadish (c'est-à-dire « Quelle approche sert le mieux la justice sociale? »), tant cette notion de justice sociale est chargée de valeurs, mais pour l'instant la question implicite est sans objet, tant que nous n'avons pas beaucoup plus d'évaluateurs qu'aujourd'hui qui fonctionnent en tant que critiques sociaux.

Conclusions et souhaits

[...]

Les évaluatrices et les évaluateurs conçoivent-elles et ils des évaluations plus éthiques qu'il y a 25 ans? Sont-elles et ils plus compétent-e-s pour relever les défis éthiques qui se posent dans leur travail? La réponse

courte, et probablement la plus longue aussi, c'est que nous ne le savons pas. Mais on peut dire sans risque que les évaluateurs et les évaluatrices sont mieux équipé-e-s pour faire ces choses qu'ils et elles ne l'étaient auparavant, en supposant qu'ils et elles restent au courant de la littérature du domaine. Notre inventaire d'analyses éthiques fondées sur des cas, dans lequel les principes directeurs et les normes d'évaluation des programmes sont appliqués, est substantiel et grandissant. Les articles dans lesquels les questions éthiques sont référencées et explorées sont devenus monnaie courante. Des recherches empiriques permettant d'identifier et d'explorer les problèmes d'éthique rencontrés par les évaluatrices et les évaluateurs au cours des différentes phases d'une évaluation (par exemple, conflits d'intérêts, pressions visant à déformer les résultats et suppression des résultats) sont maintenant disponibles (Morris et Clark, 2009; Morris et Cohn, 1993; Morris et Jacobs, 2000; Turner, 2003).

Un thème crucial qui se dégage de toute cette littérature est l'importance de l'étape de la négociation initiale du contrat d'évaluation pour établir le ton éthique de celle-ci. La base des connaissances dans le champ de l'évaluation s'est développée à un point tel que les évaluatrices et les évaluateurs peuvent compter davantage que sur la simple spéculation lorsqu'elles et ils envisagent les questions éthiques clés qui sont susceptibles d'être abordées dans le cadre d'un projet d'évaluation particulier, y compris celles qui impliquent le rôle de l'évaluateur ou de l'évaluatrice. Le fait de traiter ces questions avec les parties prenantes au début d'un projet peut aider à prévenir de graves problèmes plus tard dans l'évaluation. Tous les problèmes éthiques peuvent-ils être anticipés de cette manière? Non, mais beaucoup peuvent l'être. Les conflits qui se manifestent aux étapes ultérieures d'une évaluation (par exemple, les différends au sujet de la diffusion d'un rapport final) ont souvent leur origine dans des hypothèses ou des malentendus qui n'ont pas été examinés dans les conversations initiales. Il serait vain de faire des prédictions sur l'avenir de l'éthique dans le domaine de l'évaluation dans son ensemble; c'est pourquoi, dans cet essai, je préfère partager des

souhaits plutôt que de faire des prévisions. Il est plus facile d'anticiper le terrain éthique qu'un évaluateur ou qu'une évaluatrice devra parcourir dans un projet précis. Les évaluateurs et les évaluatrices sont bien mieux à même d'anticiper aujourd'hui qu'ils ne l'étaient en 1986, et c'est une bonne chose, même si la discussion se poursuit sur la façon dont les évaluateurs et les évaluatrices pourraient au mieux remplir les responsabilités du champ de l'évaluation à l'égard de l'intérêt public.

Bibliographie

Alexander, Leslie B. et Kenneth A. Richman. 2008. « Ethical dilemmas in evaluations using indigenous research workers ». *American Journal of Evaluation* 29(1) : 73-85.

Altschuld, James W. 1999a. « The case for a voluntary system for credentialing evaluators ». *American Journal of Evaluation* 20(3) : 507-17.

Altschuld, James W. 1999b. « The certification of evaluators: Highlights from a report submitted to the board of directors of the American Evaluation Association ». *American Journal of Evaluation* 20(3) : 481-93.

American Evaluation Association. 1995. « Guiding principles for evaluators ». in *Guiding principles for evaluators. New directions for evaluation.* N°66, édité par W. R. Shadish, D. L. Newman, M. A. Scheirer et C. Wye. San Francisco: Jossey-Bass, p. 19-26.

American Evaluation Association. 2004. « Guiding principles for evaluators ». En ligne : http://www.eval.org/Publications/GuidingPrinciples.asp

Azzam, Tarek. 2010. « Evaluator responsiveness to stakeholders ». *American Journal of Evaluation* 31(1) : 45-65. doi : https://doi.org/10.1177%2F1098214009354917.

Balch, George I., et Donna M. Mertens. 1999. « Focus group design and dynamics: Lessons from deaf and hard of hearing participants ». *American Journal of Evaluation* 20(2) : 265-77.

Barkdoll, Gerald L. 1992. « Strong medicine and unintended consequences ». *Evaluation Practice* 13(1) : 53-57.

Bickman, Leonard. 1999. « AEA, bold or timid? » *American Journal of Evaluation* 20(3) : 519-20. doi : https://doi.org/10.1177%2F109821409902000310

Birman, Dina. 2007. « Sins of omission and commission: To proceed, decline, or alter? » *American Journal of Evaluation* 28(1) : 79-85. doi : https://doi.org/10.1177%2F1098214006298059

Botcheva, Luba, Johanna Shih et Lynne C. Huffman. 2009. « Emphasizing cultural competence in evaluation: A process-oriented approach ». *American Journal of Evaluation* 30 : 176-88.

Botcheva, Luba, Catherine R. White et Lynne C. Huffman. 2002. « Learning cultures and outcomes measurement practices in community agencies ». *American Journal of Evaluation* 23(4) : 421-34. doi : https://doi.org/10.1177%2F109821400202300404

Campbell, Rebecca, Adrienne E. Adams et Debra Patterson. 2008. « Methodological challenges of collecting data from traumatized clients/consumers: A comparison of three methods ». *American Journal of Evaluation* 29 : 369-81. doi : https://doi.org/10.1177%2F1098214008320736

Chelimsky, Eleanor. 1995. « The political environment of evaluation and what it means for the development of the field ». *Evaluation Practice* 16(3) : 215-25. doi : https://doi.org/10.1177%2F109821409501600301

Chelimsky, Eleanor. 2008. « A clash of cultures : Improving the “fit” between evaluative independence and the political requirements of a democratic society ». *American Journal of Evaluation* 29(4) : 400-415. doi : https://doi.org/10.1177%2F1098214008324465

Chelimsky, Eleanor, David Cordray, et Lois-Ellin Datta. 1989. « Federal evaluation: The pendulum has swung too far ». *Evaluation Practice* 10 : 25-30.

Clayson, Zoe C., Xóchitl Castaneda, Emma Sanchez et Claire Brindis. 2002. « Unequal power-Changing landscapes: Negotiations between evaluation stakeholders in Latino communities ». *American Journal of Evaluation* 23(1) : 33-44. doi : https://doi.org/10.1177%2F109821400202300104

Cooksy, Leslie J. 2005. « The complexity of the IRB process: Some of the things you wanted to know about IRBs but were afraid to ask ». *American Journal of Evaluation* 26(3) : 352-61. doi : https://doi.org/10.1177%2F109821400502600307

Cooksy, Leslie J. 2007. « Should we call the whole thing off? » *American Journal of Evaluation* 28 : 77-78.

Cousins, J. Bradley. 2004. « Commentary: Minimizing evaluation misuse as principled practice ». *American Journal of Evaluation* 25(3) : 391-97. doi : https://doi.org/10.1177%2F109821400402500311

Datta, Lois-Ellin. 1999. « CIRCE’s demonstration of a close-to-ideal evaluation in a less-than-ideal world ». *American Journal of Evaluation* 20(2) : 345-54. doi : https://doi.org/10.1177%2F109821409902000215

Datta, Lois-Ellin. 2000. « Seriously seeking fairness: Strategies for crafting non-partisan evaluations in a partisan world ». *American Journal of Evaluation* 21(1) : 1-14. doi : https://doi.org/10.1177%2F109821400002100101

Datta, Lois-Ellin. 2002. « The case of the uncertain bridge ». *American Journal of Evaluation* 23(2) : 187-206. doi : https://doi.org/10.1177%2F109821400202300207

Dewey, Jennifer D., Bianca E. Montrosse, Daniela C. Schröter, Carolyn D. Sullins et John R. Mattox. 2008. « Evaluator competencies: What's taught versus what's sought ». *American Journal of Evaluation* 29 : 268-87.

Eddy, Rebecca M., et Tiffany Berry. 2009. « The evaluator's role in recommending program closure: A model for decision making and professional responsibility ». *American Journal of Evaluation* 30(3) : 363-76. doi : https://doi.org/10.1177%2F1098214009339931

English, Brian. 1997. « Conducting ethical evaluations with disadvantaged and minority target groups ». *Evaluation Practice* 18(1) : 49-54. doi : https://doi.org/10.1177%2F109821409701800105

Ferris, Lori E. 2000. « Legal and ethical issues in evaluating abortion services ». *American Journal of Evaluation* 21(3) : 329-40. doi : https://doi.org/10.1177%2F109821400002100304

Fetterman, David M. 1994. « Empowerment evaluation ». *Evaluation Practice* 15(1) : 1-15. doi : https://doi.org/10.1177%2F109821409401500101

Fetterman, David M., et Abraham Wandersman. 2007. « Empowerment evaluation: Yesterday, today, and tomorrow ». *American Journal of Evaluation* 28(2) : 179-98. doi : http://dx.doi.org/10.1177/1098214007301350

Fitzpatrick, Jody L. 2007. « Evaluation of the Fun with Books program: An interview with Katrina Bledsoe ». *American Journal of Evaluation* 28 : 522-35.

Fleischer, Dreolin N., et Christina A. Christie. 2009. « Evaluation use: Results from a survey of U.S. American Evaluation Association members ». *American Journal of Evaluation* 30(2) : 158-75. doi : https://doi.org/10.1177%2F1098214008331009

Friedman, Victor J., Jay Rothman et Bill Withers. 2006. « The power of why: Engaging the goal paradox in program evaluation ». *American Journal of Evaluation* 27 : 201-18.

Gill, Joan C. 1999. « Invisible ubiquity: The surprising relevance of disability issues in evaluation ». *American Journal of Evaluation* 20 : 279-87.

Greene, Jennifer C., et Jin-Hee Lee. 2006. « Quieting educational reform… with educational reform ». *American Journal of Evaluation* 27(3) : 337-52. doi : 10.1177/1098214006291103

Hedrick, Terry E. 1988. « The interaction of politics and evaluation ». *Evaluation Practice* 9(3) : 5-14.

Hennessy, Michael, et Martin J. Sullivan. 1989. « Good organizational reasons for bad evaluation research ». *Evaluation Practice* 10 : 41-50.

Hennessy, Michael, et Martin J. Sullivan. 1990. « Academic vs client critique: A response to Lipsey ». *Evaluation Practice* 11 : 165-66.

Henry, Gary T., et Melvin M. Mark. 2003. « Toward an agenda for research on evaluation ». in *The practice-theory relationship in evaluation. New directions for evaluation* N°97, édité par C. A. Christie. San Francisco: Jossey-Bass, p. 69-80.

Hopson, Rodney K. 1999. « Minority issues in evaluation revisited: Reconceptualizing and creating opportunities for institutional change ». *American Journal of Evaluation* 20(3) : 445-51.

Hopson, Rodney K. 2001. « Global and local conversations on culture, diversity, and social justice in evaluation: Issues to consider in a 9/11 era ». *American Journal of Evaluation* 22 : 375-80.

House, Ernest R. 1995. « Principled evaluation: A critique of the AEA guiding principles ». in *Guiding principles for evaluators. New directions for evaluation* N°66, édité par W. R. Shadish, D. Newman, M. A. Scheirer et C. Wye. San Francisco: Jossey-Bass, p. 27-34.

House, Ernest R. 1999. « Evaluation and people of color: A response to Professor Stanfield ». *American Journal of Evaluation* 20(3) : 433-35.

House, Ernest R., et Kenneth R. Howe. 1999. *Values in Evaluation and Social Research*. 1re éd. Thousand Oaks: Sage Publications.

Howell, Embry M., et Alshadye Yemane. 2006. « An assessment of evaluation designs : Case studies of 12 large federal evaluations ». *American Journal of Evaluation* 27 : 219-36.

Hurwith, Rosalind, et Martin Sweeney. 1995. « The use of the visual image in a variety of Australasian evaluations ». *Evaluation Practice* 16(2) : 153-64.

Iriti, Jennifer E., William E. Bickel, et Catherine A. Nelson. 2005. « Using recommendations in evaluation : A decision-making framework for evaluators ». *American Journal of Evaluation* 26(4) : 464-79.

Jaycox, Lisa *et al*. 2006. « Challenges in the evaluation and implementation of school-based prevention and intervention programs on sensitive topics ». *American Journal of Evaluation* 27 : 320-36.

Jewiss, Jennifer, et Kelly Clark-Keefe. 2007. « On a personal note: Practical pedagogical activities to foster the development of "reflective practitioners" ». *American Journal of Evaluation* 28(3) : 334-47. doi: https://doi.org/10.1177%2F1098214007304130

Johnson, Knowlton, Denise Bryant, Dward Rockwell, *et al.* 1999. « Obtaining active parental consent for evaluation research: A case study ». *American Journal of Evaluation* 20(2) : 239-49. doi : https://doi.org/10.1177%2F109821409902000206

Johnson, Timothy P., Ananda Mitra, Ramona Newman et John Horm. 1993. « Problems of definition in sampling special populations: The case of homeless persons ». *Evaluation Practice* 14(2) : 119-26. doi : https://doi.org/10.1177%2F109821409301400201

Joint Committee on Standards for Educational Evaluation. 1994. *The program evaluation standards: How to assess evaluations of educational programs*. 2e éd. Thousand Oaks: Sage Publications.

Kirkhart, Karen E. 1995. « Seeking multicultural validity: A postcard from the road ». *Evaluation Practice* 16(1) : 1-12.

Lackey, Jill F., D. Paul Moberg, et Matt Balistrieri. 1997. « By whose standards? Reflections on empowerment evaluation and grassroots groups ». *Evaluation Practice* 18 : 137-46.

Leakey, Tricia, Kevin B. Lunde, Karin Koga, et Karen Glanz. 2004. « Written parental consent and the use of incentives in a youth smoking prevention trial: A case study from Project SPLASH ». *American Journal of Evaluation* 25(4) : 509-23.

Lee, Jin-Hee, et Daniel J. Walsh. 2004. « Quality in early childhood programs: Reflections from program evaluation practices ». *American Journal of Evaluation* 25(3) : 351-73. doi : https://doi.org/10.1177%2F109821400402500306

Lipsey, Mark W. 1988. « Practice and malpractice in evaluation research ». *Evaluation Practice* 9 : 5-24.

Lipsey, Mark W. 1990. « The experimental paradigm ». *Evaluation Practice* 11(1) : 99-100. doi : https://doi.org/10.1177%2F109821409001100118

Mabry, Linda. 1999. « Circumstantial ethics ». *American Journal of Evaluation* 20(2) : 199-212. doi : https://doi.org/10.1177%2F109821409902000203

Mangano. 1991. « Using the media to disseminate evaluation results ». *Evaluation Practice* 12(2) : 115-20.

Mertens, Donna M. 1999. « Inclusive evaluation: Implications of transformative theory for evaluation ». *American Journal of Evaluation* 20(1) : 1-14. doi : https://doi.org/10.1177%2F109821409902000102

Mertens, Donna M. 2001. « Inclusivity and transformation: Evaluation in 2010 ». *American Journal of Evaluation* 22(3) : 367-74. doi : https://doi.org/10.1177%2F109821400102200313

Mertens, Donna M. 2007. « Transformative considerations: Inclusion and social justice ». *American Journal of Evaluation* 28(1) : 86-90. doi : https://doi.org/10.1177%2F1098214006298058

Morabito, Stephen M. 2002. « Evaluator roles and strategies for expanding evaluation process influence ». *American Journal of Evaluation* 23(3) : 321-30.

Morris. 1998. « The Downside report ». *American Journal of Evaluation* 19(1) : 73.

Morris, Michael. 2003. « Ethical considerations in evaluation ». in *International handbook of educational evaluation: Part one. Perspectives*, édité par T. Kellaghan et D. L. Stufflebeam. Dordrecht: Kluwer Academic, p. 303-28.

Morris, Michael, éd. 2008. *Evaluation ethics for best practice: Cases and commentaries*. New York: Guilford Press.

Morris, Michael, et B. Clark. 2009. « You want me to do WHAT? Evaluators and the pressure to misrepresent findings ». Orlando, FL, USA.

Morris, Michael, et Robin Cohn. 1993. « Program evaluators and ethical challenges: A national survey ». *Evaluation Review* 17(6) : 621-42. doi : https://doi.org/10.1177%2F0193841X9301700603

Morris, Michael, et Lynette R. Jacobs. 2000. « You got a problem with that? Exploring evaluators' disagreements about ethics ». *Evaluation Review* 24(4) : 384-406.

Newman, Dianna L., et Robert D. Brown. 1996. *Applied ethics for program evaluation*. Thousand Oaks: Sage Publications.

Patton, Michael Q. 1988. « Six honest serving men ». *Studies in Educational Evaluation* 14(3) : 301-30.

Patton, Michael Q. 1999. « Some framing questions about racism and evaluation: Thoughts stimulated by Professor Stanfield's "Slipping through the front door" ». *American Journal of Evaluation* 20(3) : 437-43. doi : https://doi.org/10.1177%2F109821409902000303

Penuel, William R., Willow Sussex, Christine Korbak, et Christopher Hoadley. 2006. « Investigating the potential of network analysis in educational evaluation ». *American Journal of Evaluation* 27(4) : 437-51. doi : https://doi.org/10.1177%2F1098214006294307

Preskill, Hallie, et Valerie Caracelli. 1997. « Current and developing conceptions of use: Evaluation Use TIG survey results ». *Evaluation Practice* 18(1) : 209-25. doi : https://doi.org/10.1177%2F109821409701800122

Renger, Ralph. 2006. « Consequences to federal programs when the logic-modeling process is not followed with fidelity ». *American Journal of Evaluation* 27(4) : 452-63. doi : https://doi.org/10.1177%2F1098214006293666

Renger, Ralph, Adriana Cimetta, Sydney Pettygrove et Seumas Rogan. 2002. « Geographic Information Systems (GIS) as an evaluation tool ». *American Journal of Evaluation* 23(4) : 469-79. doi : https://doi.org/10.1177%2F109821400202300407

Rossi, Peter H. 1995. « Doing good and getting it right ». in *Guiding principles for evaluators. New directions for evaluation.* N°66, édité par W. R. Shadish, D. L. Newman, M. A. Scheirer et C. Wye. San Francisco: Jossey-Bass, p. 55-59.

Ryan, Katherine. 2002. « Shaping educational accountability systems ». *American Journal of Evaluation* 23(4) : 453-68. doi : https://doi.org/10.1177%2F109821400202300406

Schwandt, Thomas A. 1992. « Better living through evaluation? Images of progress shaping evaluation practice ». *Evaluation Practice* 13(2) : 135-44. doi : https://doi.org/10.1177%2F109821409201300206

Schwandt, Thomas A. 2008. « Educating for intelligent belief in evaluation ». *American Journal of Evaluation* 29(2) : 139-50. doi : https://doi.org/10.1177%2F1098214008316889

Scriven, Michael. 1997. « Empowerment evaluation examined ». *Evaluation Practice* 18(1) : 165-75. doi : https://doi.org/10.1177%2F109821409701800115

Scriven, Michael. 2005. « Review of Empowerment evaluation principles in practice ». *American Journal of Evaluation* 26 : 415-17.

Shadish, William R. 1994. « Need-based evaluation: Good evaluation and what you need to know to do it ». *Evaluation Practice* 15(3) : 347-58. doi : https://doi.org/10.1016/0886-1633(94)90029-9

Shaw, Ian, et Alex Faulkner. 2006. « Practitioner evaluation at work ». *American Journal of Evaluation* 27(1) : 44-63. doi : https://doi.org/10.1177%2F1098214005284968

Simons, Helen. 2006. « Ethics in evaluation ». in *The SAGE handbook of evaluation*, édité par I. F. Shaw, J. C. Greene et M. M. Mark. Thousand Oaks: Sage Publications, p. 243-65.

Slaughter, Helen B. 1991. « The participation of cultural informants on bilingual and cross-cultural evaluation teams ». *Evaluation Practice* 12(2) : 149-57. doi : https://doi.org/10.1177%2F109821409101200205

Smith, M. F. 1999. « Should AEA begin a process for restricting membership in the profession of evaluation? » *American Journal of Evaluation* 20(3) : 521-31. doi : https://doi.org/10.1177%2F109821409902000311

Smith, Nick L. 1998. « Professional reasons for declining an evaluation contract ». *American Journal of Evaluation* 19(2) : 177-90.

Smith, Nick L. 2002. « An analysis of ethical challenges in evaluation ». *American Journal of Evaluation* 23(2) : 199-205.

Spicer, Rebecca, Valerie Nelkin, Ted Miller, et Les Becker. 2004. « Using corporate data in workplace program evaluation ». *American Journal of Evaluation* 25 : 109-19.

Stanfield, John H. 1999. « Slipping through the front door: Relevant social scientific evaluation in the people-of-color century ». *American Journal of Evaluation* 20 : 415-31.

Stevahn, Laurie, Jean A. King, Gail Ghere, et Jane Minnema. 2005. « Establishing essential competencies for program evaluators ». *American Journal of Evaluation* 26(1) : 43-59.

Stevens, Carla J., et Micah Dial. 1994. *Preventing the misuse of evaluation. New directions for program evaluation*, N°64. San Francisco: Jossey-Bass.

Stufflebeam, Daniel L. 1994. « Empowerment evaluation, objectivist evaluation, and evaluation standards: Where the future of evaluation should not go and where it needs to go ». *Evaluation Practice* 15(3) : 321-38. doi : https://doi.org/10.1177%2F109821409401500313

Taut, Sandy M., et Marvin C. Alkin. 2003. « Program staff perceptions of barriers to evaluation implementation ». *American Journal of Evaluation* 24(2) : 213-26.

Torres, Rosalie T., Hallie S. Preskill, et Mary E. Piontek. 1997. « Communicating and reporting : Practices and concerns of internal and external evaluators ». *Evaluation Practice* 18(2) : 105-25. doi : https://doi.org/10.1177%2F109821409701800203

Turner, D. 2003. « Evaluation ethics and quality: Results of a survey of Australasian Evaluation Society members ».

Walker, Robert, Lesley Hoggart, et Gayle Hamilton. 2008. « Random assignment and informed consent: A case study of multiple perspectives ». *American Journal of Evaluation* 29(2) : 156-74. doi : https://doi.org/10.1177%2F1098214008317206

Wandersman, Abraham, et Jessica Snell-Johns. 2005. « Empowerment evaluation: Clarity, dialogue, and growth ». *American Journal of Evaluation* 26(3):421-28. doi : https://doi.org/10.1177%2F1098214005278774

Wandersman, Abraham, Jessica Snell-Johns, *et al.* 2005. « The principles of empowerment evaluation ». in *Empowerment evaluation principles in practice*, édité par D. Fetterman et A. Wandersman. New York: The Guilford Press, p. 27-41.

Williams, David D. 1991. « Lessons learned: Introducing accreditation and program evaluations to a teachers college in Indonesia ». *Evaluation Practice* 12(1) : 23-31.

Wolf, Amanda, David Turne et Kathleen Toms. 2009. « Ethical perspectives in program evaluation ». in *The handbook of social research ethics*, édité par D. M. Mertens et P. E. Ginsberg. Los Angeles: Sage Publications, p. 170-83.

Worthen, Blaine R. 1999. « Critical challenges confronting certification of evaluators ». *American Journal of Evaluation* 20(3) : 533-55. doi: https://doi.org/10.1177%2F109821409902000312

Yarbrough, Donald B., Lyn M. Shulha, Rodney K. Hopson et Flora A. Caruthers. 2011. *The program evaluation standards: A guide for evaluators and evaluation users*. Los Angeles: Sage Publications.

3. La vision démocratique délibérative

ERNEST R. HOUSE ET KENNETH R. HOWE

[Traduit de House, Ernest R., Kenneth R. Howe. 1999, *Values in Evaluation and Social Research*, Thousand Oaks: Sage, chapitre 6 "The deliberative democratic view" p.91-103 (extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Au cours des années 1980, la Division de l'évaluation et de la méthodologie des programmes (*Program evaluation and methodology division*, ou PEMD) du *General Accounting Office* des États-Unis était l'unité d'évaluation la plus réputée de Washington. Sa directrice, Eleanor Chelimsky (1998), a fourni un précieux résumé de ce qu'elle a appris au cours de ses années à la tête de ce bureau. L'une de ses conclusions est que les conditions politiques spécifiques ont de fortes répercussions sur la façon dont les évaluations sont effectuées. Elle prend aussi cette position très ferme au sujet du plaidoyer :

> Dans un environnement politique, il n'est pas nécessaire de faire entendre une autre voix revendicative, mais plutôt de rendre disponible pour un usage public une information robuste, honnête, et sans parti pris pour quelque cause que ce soit. Les responsables politiques du Congrès s'attendent à ce que les évaluateurs et les évaluatrices jouent exactement un tel rôle et apportent précisément ce genre d'informations…Pourtant, nous avons vu récemment des tentatives visant à justifier une posture revendicative de l'évaluateur ou de l'évaluatrice, et certaines théories appuient cette idée…Notre expérience au PEMD est que

> la défense d'intérêts de quelque nature que ce soit détruit la crédibilité de l'évaluateur ou de l'évaluatrice et n'a pas sa place dans l'évaluation (p.40).

Elle note en même temps que le Congrès remet rarement en question de façon sérieuse les programmes du Département de la Défense. Elle a trouvé cela particulièrement vrai pour les questions relatives à la guerre chimique[1]. En 1981, lorsque Chelimsky entreprend des études sur les programmes de guerre chimique, elle découvre qu'il y a deux ensembles de publications. Le premier était classifié, favorable aux armes chimiques, et présenté par le Pentagone d'une manière unilatérale au Congrès. L'autre était critique, pro-paix, publique, et les responsables politiques du Congrès n'envisageaient même pas de le prendre en compte.

En découvrant cette situation, ses services ont réalisé une synthèse de toute la littérature, dit-elle, « qui a eu un effet électrisant sur les membres du Congrès qui affrontaient pour la première fois certains faits » (p.43). Ce document initial a donné lieu à davantage d'évaluations, de publicité et, à terme, a contribué aux accords internationaux sur les armes chimiques – une évaluation réussie selon presque toutes les normes en vigueur.

Ce travail sur la guerre chimique reposait sur l'analyse des tendances partisanes de la recherche existante, la compréhension des fondements politiques du programme et de l'évaluation, et la tentative d'« intégrer des valeurs contradictoires » à l'évaluation – ce que Chelimsky recommande pour toutes ces études. C'est une approche très intelligente, nous semble-t-il.

Notre question est : Quel cadre l'a amenée à mener l'étude de cette façon? Pourquoi a-t-elle posé des questions sérieuses sur les programmes du Pentagone alors que le Congrès ne l'a pas fait? Aucun groupe de parties

1. NdT : Cette expérience a été racontée en français dans Chelimsky, E. (1985). "L'évaluation de programmes aux États-Unis". *Politiques et management public*, 3(2), p. 199-214.

prenantes ne l'a incitée à le faire. Le Pentagone a poussé ses propres informations, et les colombes anti-chimiques les leurs. Chelimsky devait avoir un cadre, aussi intuitif que cela ait pu être, pour la guider sur ce qu'il fallait faire.

Nous ne savons pas quel cadre elle a réellement utilisé, mais nous pensons qu'un cadre qui pourrait produire des résultats similaires prendrait la forme suivante : Inclure les valeurs contradictoires et les groupes de parties prenantes dans l'étude. S'assurer que tous les points de vue importants sont suffisamment inclus et représentés. Rassembler des points de vue contradictoires afin qu'il y ait des délibérations et un dialogue à leur sujet entre les parties concernées. Non seulement s'assurer qu'il y a suffisamment de place pour le dialogue pour résoudre les revendications contradictoires, mais aussi aider les décideuses et les décideurs et les médias à faire justice de ces revendications en triant les bonnes et les mauvaises informations. Mettre sur le devant de la scène les intérêts des bénéficiaires présumé-e-s s'ils et elles sont négligé-e-s.

Toute cette analyse et cette interprétation exigent de nombreux jugements et décisions de la part des évaluateurs et des évaluatrices quant à savoir ce qui est pertinent, ce qui est important, ce qui est une bonne information et ce qui est mauvais, comment gérer les délibérations des décideurs, comment gérer les médias, quelles sont les implications politiques, etc. Les évaluatrices et les évaluateurs sont inévitablement impliqué-e-s dans les constats qu'elles et ils font, même si elles et ils ne formulent pas elles/eux-mêmes les conclusions de l'étude. Leur empreinte intellectuelle est partout.

Il y a plusieurs constats à faire ici. Le premier est qu'un cadre est nécessaire pour guider l'évaluation, même s'il est implicite. Deuxièmement, ce cadre est une combinaison de faits et de valeurs. Les valeurs relatives à la guerre chimique des différents groupes impliqués ont été un élément important de l'évaluation. Les faits et les valeurs ont

été réunis [...]. De plus, l'évaluation sur la guerre chimique par Chelimsky est guidée par une conception particulière du rôle de l'évaluation dans les politiques publiques.

Peut-on alors parler d'une posture revendicative de la part des évaluateurs et des évaluatrices? Nous dirions que non, même si le travail mené est très chargé de valeurs et intègre une part considérable de jugement de la part des évaluateurs et des évaluatrices. Il ne s'agit pas de défendre le Pentagone ou les colombes au début de l'étude et de ne défendre qu'un seul côté ou l'autre. Après tout, si le Congrès est si fortement orienté vers le Pentagone, il serait politiquement logique de rester de leur bon côté, parce qu'ils sont les clients. C'est probablement ce que les évaluatrices et les évaluateurs qui cherchent à répondre aux besoins de leurs clients et leurs clientes (*client-oriented evaluators*) auraient fait. Ou il se peut qu'ils et elles auraient construit des résumés de valeur (*value summaries*) tels que ceux approuvés par Shadish, Cook, et Leviton (1995) : « Si vous êtes en faveur des armes chimiques, X est l'action à engager, mais si vous y êtes opposés, Y est l'action à prendre », et les auraient fournis aux décideurs et aux décideuses.

Mais les évaluateurs et les évaluatrices ont fait quelque chose de plus défendable – ils et elles ont impliqué toutes les parties dans l'étude et ont évalué la qualité des preuves de chaque côté en les comparant aux assertions critiques de l'autre côté. C'était ce qu'il fallait faire.

La conduite de cette étude est conforme au type de théorie de l'évaluation que nous voulons soutenir. L'approche démocratique délibérative que nous proposons intègre trois critères généraux pour que les évaluations soient correctement équilibrées sur les plans des valeurs, des parties prenantes et de la politique. Premièrement, l'étude devrait être inclusive de manière à représenter tous les points de vue, intérêts, valeurs, intervenants et intervenantes pertinent-e-s. Aucun élément important ne doit être omis. Dans le cas de la guerre chimique, les points de vue

critiques des programmes de guerre chimique avaient été omis à l'origine et seuls les points de vue favorables au Pentagone avaient été inclus, ce qui a faussé les conclusions des études précédentes.

Deuxièmement, il devrait y avoir un dialogue suffisant avec les groupes concernés afin que les points de vue soient dûment et authentiquement représentés. Il n'est pas toujours facile de représenter ces points de vue de façon fidèle, mais c'est souvent essentiel. « Le fait de prêter attention à ce que pensent les bénéficiaires d'un programme est une caractéristique essentielle d'une étude crédible et n'a rien à voir avec la défense de ces bénéficiaires » (Chelimsky, 1998 : 47). De nombreuses études ont été menées sans tenir compte des intérêts des principaux et principales bénéficiaires (ou des victimes). Dans ce cas, les victimes potentielles de la guerre chimique ne peuvent guère être présentes. Quelqu'un doit représenter leurs intérêts. Sans doute, inclure les parties prenantes et leur parler dans la mesure du possible n'est pas un plaidoyer de l'avis de Chelimsky.

Troisièmement, il doit y avoir un niveau de délibération suffisante pour arriver à des constats de qualité. Dans ce cas, la délibération fut longue et productive, et a inclus les évaluateurs/-trices, les décideurs/-euses, et *in fine* les médias. La délibération peut inclure des moyens de protéger les évaluateurs et les évaluatrices et d'autres des pressions de puissantes parties prenantes, qui peuvent sérieusement entraver les discussions, comme le note Chelimsky. Une bonne délibération ne peut pas simplement être une discussion libre entre parties prenantes. Si c'est le cas, les parties prenantes puissantes gagnent.

La conception et la gestion de tout cela impliquent un effort considérable de jugement de la part des évaluateurs et des évaluatrices. Les évaluateurs et les évaluatrices peuvent être guidé-e-s par l'intuition, comme Chelimsky et ses collègues ont semblé l'être, ou ils et elles peuvent compter sur quelque chose de plus explicite. En fait, Chelimsky avance une conception particulière de l'intérêt public, c'est-à-dire que l'évaluation doit être jugée en fonction « de son succès à fournir des

informations objectives dans l'intérêt public » (p.52). Et elle va plus loin : « Je suppose que le risque beaucoup plus grand pour notre domaine n'est pas le manque d'utilisation pour les bonnes raisons, mais plutôt le déclin de la capacité ou de la volonté de remettre en question le sens commun, ce qui est notre tâche la plus importante et la meilleure justification de notre travail » (p.51).

Ce faisant, n'est-elle pas en train de faire du plaidoyer pour sa conception particulière de l'intérêt public et du rôle de l'évaluation dans ce domaine? Si ce n'est pas le cas, en quoi sa vision diffère-t-elle d'une posture revendicative? La défense d'un point de vue implique de prendre les points de vue ou les intérêts d'un groupe et de toujours les défendre contre les autres, indépendamment des résultats de l'évaluation. Par exemple, Chelimsky et ses collègues auraient pu prendre soit le point de vue du Pentagone, soit celui des colombes sans équilibrer les deux. Ce serait une sorte de plaidoyer. Elle n'a pas fait cela.

D'un autre côté, si le plaidoyer signifie utiliser ou approuver des cadres ou des valeurs particuliers, Chelimsky pourrait être accusée de défendre sa conception particulière de l'intérêt public, avec laquelle tout le monde ne serait pas d'accord. Elle dit que tous les évaluateurs devraient procéder à des évaluations en tenant compte de l'intérêt public. Elle pourrait faire du plaidoyer en ce sens qu'elle soutient le recours à un cadre général. En fait, nous croyons que tous les évaluateurs et toutes les évaluatrices doivent adopter une certaine conception de l'intérêt public et de la démocratie, même si ces conceptions sont implicites.

En *ce* sens, les évaluatrices et les évaluateurs devraient faire du plaidoyer – en faveur de la démocratie et de l'intérêt public. La démocratie aspire à intégrer tous les intérêts légitimes. À notre avis, l'intérêt public n'est pas statique et, souvent, il n'est pas identifiable au départ, mais il émerge (ou devrait) de manière appropriée des processus démocratiques restreints dans lesquels l'évaluation joue un rôle. C'est précisément parce que les évaluateurs et les évaluatrices *devraient être des défenseurs et les défenseuses* de la démocratie et de l'intérêt public, qu'ils et elles *ne*

devraient pas défendre des groupes spécifiques de parties prenantes dont les intérêts perçus résistent aux éléments de preuves et sont promus quoi qu'il arrive (Greene, 1997, utilise le terme *plaidoyer* dans un sens et Chelimsky, 1998, dans un autre, donc malheureusement leurs conceptions ne dialoguent pas). Les évaluateurs et les évaluatrices ne devraient pas non plus jouer le rôle de facilitateurs et de facilitatrices neutres parmi les défenseurs et les défenseuses de « résumés de valeur » concurrents ou des « constructions » des parties prenantes, de notre point de vue.

En quoi ce cas de la guerre chimique diffère-t-il de l'évaluation des programmes sociaux? Très peu. Dans leur évaluation des services de santé sur la côte du golfe du Texas, Madison et Martinez (1994) ont identifié comme principales parties prenantes les bénéficiaires des services (des Afro-Américain-e-s âgé-e-s) et les fournisseurs et les fournisseuses de services (surtout des médecins et infirmières blancs), ainsi que des représentant-e-s de groupes de défense des intérêts afro-américains. Chaque groupe avait un point de vue différent, les personnes âgées déclarant que les services n'étaient pas suffisamment accessibles et les professionnel-le-s de la santé déclarant que les personnes âgées n'étaient pas au courant de ces services.

Est-ce du plaidoyer que d'inclure des groupes particuliers, disons les Afro-Américain-e-s âgé-e-s dans ce cas, dans l'étude? Nous pensons qu'il ne s'agit pas d'un plaidoyer, mais plutôt d'un équilibre entre les valeurs et les intérêts de l'étude. Tous les points de vue devraient être représentés. Ce n'est pas non plus du plaidoyer que de rentrer dans cette étude en sachant que les opinions afro-américaines sont souvent exclues de ces études. C'est une histoire documentée, et les évaluateurs et les évaluatrices devraient être attentifs et attentives à de tels risques d'exclusion.

Dans une telle évaluation, on ne va pas établir pour de bon la légitimité des droits des personnes âgées afro-américaines par rapport à ceux des professionnel-le-s blanc-he-s dans la société en général. Cela dépasse le cadre de la plupart des évaluations. Les utilisateurs et les utilisatrices

doivent déterminer ce qui se passe avec ces services à l'heure actuelle, une tâche plus modeste. Le plaidoyer, dans un sens erroné signifierait que les évaluateurs et les évaluatrices entrent déjà dans l'étude avec l'idée que les Afro-Américain-e-s ont raison et que les fournisseurs et les fournisseuses de services se trompent, ou vice versa, quels que soient les faits. Ce n'est pas la position qui convient aux évaluateurs et aux évaluatrices professionnel-le-s.

Notre conception de l'intérêt public dans l'évaluation est que l'évaluation informe objectivement l'opinion publique en incluant les points de vue et les intérêts, en favorisant le dialogue et en favorisant les délibérations visant à tirer des conclusions valables. L'objectivité est assurée par l'inclusion, le dialogue et les délibérations ainsi que par l'expertise d'évaluation que l'évaluateur ou l'évaluatrice professionnel-le met à contribution. Les évaluateurs et les évaluatrices ne peuvent échapper à l'attachement à une certaine notion de démocratie. La question est de savoir dans quelle mesure cette notion est explicite et défendable.

Dans le reste de ce chapitre, nous détaillons cette approche démocratique délibérative, l'opposons à d'autres points de vue et esquissons son lien avec la théorie politique. Cette approche n'est pas un modèle d'évaluation en ce sens qu'il s'agit d'un cadre permettant de déterminer si une évaluation est impartiale et objective du point de vue des allégations de valeur. De même que les évaluations peuvent être faussées par une mauvaise collecte de données et des erreurs d'omission ou de commission, elles peuvent aussi être faussées quand les mauvaises valeurs, parties prenantes ou intérêts y sont intégrés. Ce cadre milite contre de tels biais.

L'évaluation démocratique délibérative

La démocratie délibérative peut être considérée comme une *véritable* démocratie, c'est-à-dire ce que la démocratie exige lorsqu'elle est correctement analysée et comprise. En un sens, l'expression « démocratie délibérative » est redondante, parce que la démocratie au sens large exige, selon nous, une délibération. Mais la redondance mérite d'être préservée afin d'éviter toute confusion au sujet de ce sur quoi nous mettons l'accent. Nous utilisons ce terme pour focaliser l'attention sur les procédures de prise de décision que la démocratie exige et pour éviter la confusion avec d'autres conceptions de la démocratie.

Notre intention est de présenter un cadre général pour juger les évaluations sur la base de leur potentiel de délibération démocratique. Trois exigences s'appliquent à l'évaluation démocratique délibérative : l'évaluation doit être inclusive, dialogique et délibérative. Nous discutons tour à tour de chacune de ces exigences, bien qu'elles ne soient pas faciles à séparer complètement les unes des autres.

L'exigence d'inclusion

La première exigence de l'évaluation démocratique délibérative est l'inclusion de tous les intérêts pertinents. Il ne serait pas juste que des évaluatrices et des évaluateurs ne fournissent des évaluations qu'aux plus puissant-e-s ou ne les vendent qu'à ceux et celles qui enchérissent le plus pour leurs propres usages, biaisant ainsi l'évaluation en direction d'intérêts particuliers. Et il ne serait pas non plus juste que les commanditaires révisent les constats, effacent les parties de l'évaluation qu'ils et elles n'aiment pas ou renforcent les constats qui conviennent à leurs conceptions intéressées. Ce sont des conditions d'utilisation que les évaluateurs et les évaluatrices ne devraient pas tolérer.

Les études évaluatives aspirent à être des représentations exactes de la réalité, et non des instruments fictifs pour promouvoir les intérêts de certain-e-s par rapport à d'autres, comme dans la publicité ou les relations publiques, le prix étant attribué à ceux et celles qui paient le service. Les intérêts de tous les groupes de parties prenantes sont essentiels et les intérêts de toutes les parties concernées devraient être représentés, comme l'exige une véritable démocratie. Si tous les intérêts pertinents ne sont pas pris en compte, le résultat n'est qu'une fausse démocratie de laquelle certains ont été exclus.

Les déséquilibres de pouvoir figurent parmi les pires menaces à l'évaluation. De tels déséquilibres sont endémiques dans la société, et on peut aisément imaginer comment ils peuvent perturber et fausser une évaluation. Les plus puissant-e-s peuvent dominer la discussion, ou ceux et celles qui sont sans pouvoir ne pas être représentés. Il doit y avoir un certain équilibre et une égalité des pouvoirs pour qu'une délibération correcte ait lieu.

Les évaluateurs et les évaluatrices doivent concevoir des évaluations de telle façon que les intérêts pertinents soient représentés et de telle façon qu'il y ait un équilibre de pouvoir entre eux, ce qui veut souvent dire représenter les intérêts de ceux et celles qui pourraient être exclu-e-s de la discussion, parce que leurs intérêts ont des chances d'être négligés en leur absence. Et, bien sûr, la délibération devrait être basée sur une discussion sur les mérites, pas sur le statut des participant-e-s.

Déterminer et pondérer les intérêts est extrêmement complexe et incertain, et induit souvent des controverses. D'abord, tous les intérêts n'ont pas la même force morale. Bhaskar (1986) distingue les intérêts qui s'attachent aux besoins, qui ont le plus de poids moral, du large éventail d'intérêts qui suit :

> Un intérêt est tout ce qui favorise la réalisation des désirs, des besoins et/ou des objectifs des agents et agentes; et un besoin est tout ce qui est (contingentement ou absolument) nécessaire à

> la survie ou au bien-être d'un-e agent-e, que celui-ci le possède ou non. La satisfaction d'un besoin, contrairement à la satisfaction d'une envie ou d'un but, ne peut jamais en soi aggraver la situation d'un individu ou un groupe (p.70).

Scriven (1991) fait valoir une distinction similaire dans le contexte spécifique de l'évaluation. Il distingue l'« évaluation de la valeur » (*value assessment*), dans laquelle les besoins, les désirs et les préférences de marché sont traités indifféremment, de l'« évaluation des besoins » bien compris. « Les besoins, dit-il, constituent la première priorité pour l'intervention.... juste parce qu'ils sont dans une certaine mesure *nécessaires*, alors que les envies sont (seulement) *désirées* » (p.241). Selon Scriven, les besoins sont associés à un « niveau d'urgence ou d'importance » que ne possèdent pas les désirs, les préférences du marché, etc.

Cela ne signifie pas qu'il soit facile, ni même toujours nécessaire, de distinguer les intérêts associés aux besoins des intérêts liés aux désirs dans la conduite des évaluations. Néanmoins, il s'agit d'une distinction à laquelle les évaluateurs et les évaluatrices devraient s'astreindre. Même si elle est floue ou controversée dans certains cas, elle n'en est pas moins tout à fait réelle. Dans de nombreux cas, il est facile de tracer la frontière – par exemple, les intérêts en matière de nourriture, de logement et de soins de santé par rapport aux intérêts en matière de retraite anticipée ou d'automobiles de luxe.

L'exigence dialogique

La deuxième exigence de l'évaluation démocratique délibérative est qu'elle soit dialogique. Ce qui complique la détermination et la hiérarchisation des intérêts, c'est que les individus et les groupes ne sont pas toujours en mesure de déterminer leurs propres intérêts lorsqu'ils

sont laissés à leurs propres moyens. Ils peuvent être dupés ou induits en erreur par les médias, par de puissants groupes d'intérêt qui suppriment ou manipulent des preuves, ou parce qu'ils et elles ne disposent pas ou ne saisissent pas les opportunités d'obtenir des informations. Les intérêts réels d'un individu ou d'un groupe ne sont pas nécessairement les mêmes que les intérêts perçus. Les intérêts réels pourraient être définis de cette manière : La politique X est dans l'intérêt de A si A, connaissant les résultats de la politique X et de la politique Y, choisirait le résultat de la politique X plutôt que celui de la politique Y. Il est essentiel de déterminer les intérêts « réels ».

Découvrir des intérêts réels est une tâche importante de l'interaction dialogique. Les évaluateurs et les évaluatrices ne peuvent pas présumer automatiquement des intérêts des parties. Ils et elles peuvent se tromper. Il est préférable d'engager activement les participant-e-s par le biais de dialogues de divers types. Il se peut que, par le dialogue et les délibérations, les parties prenantes changent d'avis quant à leurs intérêts. Après avoir examiné les conclusions et engagé dans des débats et des discussions avec d'autres, ils et elles peuvent considérer leurs intérêts comme différents de ceux avec lesquels ils et elles ont commencé.

Le fait que l'évaluation soit totalement intégrée dans le tissu social rend le dialogue essentiel. Les participant-e-s et les évaluateurs ou évaluatrices doivent identifier les problèmes réels et même les créer dans de nombreux cas. Les constats de l'évaluation apparaissent au long de ces processus. Ils n'attendent pas d'être découverts, nécessairement, mais sont forgés dans l'évaluation et les discussions sur les résultats. [...] Cela ne signifie pas que la conclusion de l'évaluation soit « irréelle » parce qu'elle est émergente et construite, pas plus qu'une voiture n'est irréelle parce qu'elle est construite.

Pour assurer que le dialogue ait lieu, les évaluateurs et les évaluatrices doivent représenter équitablement tous les intérêts, s'engager dans des processus dialogiques avec les participant-e-s et délibérer de manière approfondie sur les enjeux identifiés. Dans un certain sens, nous pouvons

imaginer d'avancer le long du continuum valeur-fait depuis des assertions portant sur les préférences et les valeurs recueillies lors du dialogue initial, puis par des délibérations fondées sur des principes démocratiques, à des déclarations évaluatives sur les faits.

Il y a un risque que les évaluateurs et les évaluatrices soient indûment influencés par un dialogue approfondi avec divers groupes de parties prenantes, une menace que Scriven (1973) a relevée il y a longtemps dans son appel à une évaluation affranchie des objectifs (*goal-free evaluation*). Nous pensons que cette menace à l'impartialité est réelle, mais le réel danger est que les évaluateurs ou les évaluatrices ne comprennent pas pleinement les positions, les opinions et les intérêts des différents groupes de parties prenantes et représentent mal ces groupes dans l'évaluation. Nous sommes donc disposés à échanger la menace d'impartialité contre la possibilité que les évaluateurs ou les évaluatrices comprennent pleinement les positions des parties prenantes en engageant un dialogue approfondi avec elles. Et la menace d'impartialité est atténuée par l'inclusion de différents groupes et les délibérations.

Dans certaines situations, il se peut qu'il y ait peu de risque d'un malentendu de la part des évaluatrices et des évaluateurs. Dans certaines évaluations de produits [comme en font les associations de consommateurs, NdT], les évaluatrices et les évaluateurs peuvent peut-être présumer des intérêts de consommateurs ou de consommatrices typiques avec un minimum de dialogue parce que le contexte de ces études peut être défini avec précision à l'avance. Toutefois, dans la plupart des évaluations de programmes et de politiques complexes, il n'est pas facile de comprendre les parties prenantes et leurs positions. Les intérêts des différents groupes peuvent être conflictuels, et plus la situation est complexe, plus le dialogue est nécessaire pour la régler. En ce sens, les évaluations de produits peuvent être plus un cas particulier d'évaluation que le cas paradigmatique. Et nous pensons que le dialogue est non seulement souhaitable, mais nécessaire dans la plupart des cas.

L'exigence délibérative

La troisième exigence des évaluations est qu'elles soient délibératives. La délibération est fondamentalement un processus cognitif, fondé sur la raison, la preuve et les principes de l'argumentation valable, dont un sous-ensemble important est les canons méthodologiques de l'évaluation. Dans de nombreux cas, l'autorité des évaluatrices et des évaluateurs, fondée sur leur expertise particulière joue un rôle essentiel dans une démocratie délibérative.

Au contraire, on considère généralement que la démocratie « émotiviste » ou préférentielle accepte telles quelles les préférences, les valeurs, les goûts et les intérêts des citoyens et citoyennes et trouve des moyens de maximiser ces intérêts. Les évaluatrices et les évaluateurs ne pourraient remettre en question ces préférences – elles seraient simplement données. Les faits se prêteraient à la détermination des spécialistes, comme dans le domaine scientifique, mais les valeurs seraient choisies et ne pourraient être traitées rationnellement. Par conséquent, le mieux que les évaluateurs et les évaluatrices pourraient faire est de satisfaire les préférences (maximiser la satisfaction des préférences), peu importe ce qu'elles sont. Ce raisonnement conduit à une conception de la démocratie dans laquelle les préférences et les valeurs ne sont pas examinées.

Nous sommes d'avis que les valeurs ne peuvent être prises telles quelles, mais qu'elles doivent faire l'objet d'un examen dans le cadre de processus rationnels. L'évaluation est une procédure de détermination des valeurs, qui sont émergentes et transformées par des processus délibératifs en constats de l'évaluation. L'évaluation sert donc une démocratie délibérative, dans laquelle les intérêts et les valeurs sont rationnellement déterminés. Une discussion et une détermination minutieuses exigent l'expertise des évaluateurs et des évaluatrices.

Certes, l'évaluation ne doit pas se substituer au vote et à d'autres procédures de décision dans une démocratie. L'évaluation est plutôt une institution qui produit des constats évaluatifs utilisés dans les processus décisionnels démocratiques. L'évaluation informe le vote et les autres procédures de décision faisant autorité dans les sociétés démocratiques, elle ne doit pas les remplacer.

Après tout, l'évaluation est inextricablement liée à la notion de choix : quels choix faut-il faire, qui fait des choix et sur quelle base? L'évaluation des programmes, des politiques et du personnel publics repose sur la notion de choix collectif et sur l'idée de tirer des conclusions sur la base de leur mérite. Au contraire, nous pouvons considérer que des individus arrivent à des conclusions en faisant l'acte individuel de pondérer et mettre en balance divers facteurs. Il s'agit d'un modèle de choix du consommateur, essentiellement un modèle de marché, dans lequel de nombreuses personnes font leurs propres choix sur la base des informations disponibles, et dans lequel le choix collectif n'est que la somme des choix individuels.

Mais la plupart des évaluations publiques ne fonctionnent pas comme cela. Les intérêts pertinents et les parties prenantes doivent être déterminés dans le cadre de l'évaluation. Et le choix du consommateur ou de la consommatrice n'est pas le même que le choix collectif découlant d'une délibération collective. La délibération collective exige une réciprocité de la conscience entre les participant-e-s et l'égalité approximative des pouvoirs si l'on veut que les participant-e-s atteignent un état dans lequel ils et elles débattent efficacement de leurs propres finalités collectives.

Une note sur l'autorité de l'évaluateur ou de l'évaluatrice en la matière : il est utile de faire la distinction entre le pouvoir et l'autorité. Les évaluateurs et les évaluatrices devraient accepter l'autorité, mais pas le pouvoir. Par exemple, A a un pouvoir sur B quand A peut affecter le comportement B d'une façon contraire aux intérêts de B. Mais A a autorité sur B lorsque B s'y conforme parce qu'A a influencé B par de bonnes

raisons liées aux intérêts de B. Il existe une délibération démocratique quand les délibérations sont des discussions sur le mérite qui mettent en jeu les intérêts de A et B ou leurs intérêts collectifs. Par conséquent, les évaluateurs et les évaluatrices ont autorité en ce sens que les gens sont persuadés par l'évaluation pour de bonnes raisons.

Les exigences d'inclusion, de dialogue et de délibération se chevauchent et se croisent de manière complexe. Par exemple, la qualité de la délibération n'est pas dissociable de la qualité du dialogue, ce qui, à son tour, influe sur la question de savoir si l'inclusion (par opposition à une participation symbolique [*tokenism*]) est obtenue. En général, les trois conditions d'inclusion, de dialogue et de délibération ne peuvent être clairement distinguées et appliquées indépendamment. Elles s'affectent et se renforcent mutuellement.

Néanmoins, les distinguer les unes des autres fournit des orientations. Si les exigences en matière d'inclusion et de dialogue sont satisfaites, mais que les délibérations ne le sont pas, tous les intérêts pertinents peuvent être représentés (provisoirement), mais ne pas être suffisamment pris en considération, ce qui donne lieu à des conclusions erronées (un problème pour les approches « constructivistes » et « postmodernistes »). Si les exigences d'inclusion et de délibération sont satisfaites, mais que le dialogue fait défaut, les intérêts et les positions peuvent être mal représentés, ce qui donne lieu à des évaluations inauthentiques fondées sur de faux intérêts et dominées par celles et ceux qui ont le plus de pouvoir [...]. Enfin, si les exigences en matière de dialogue et de délibération sont respectées, mais que tou-te-s les intervenant-e-s ne sont pas inclus-es, l'évaluation peut être accusée d'être partiale à l'égard d'intérêts particuliers.

L'évaluation démocratique délibérative est un idéal qui mérite d'être poursuivi, pas quelque chose qui peut être réalisé une fois pour toutes dans n'importe quelle étude ou pleinement pris en compte. Mais encore une fois, la collecte, l'analyse et l'interprétation de données sans biais afin d'arriver à des résultats précis n'est jamais non plus parfaite. Ce n'est pas

une raison pour les évaluateurs et les évaluatrices de cesser de faire de leur mieux. Il y a de meilleures et de pires façons de conduire les études du point de vue de la démocratie délibérative.

Bibliographie

Bhaskar, Roy. 1986. *Scientific Realism and Human Emancipation*. London: Verso.

Chelimsky, Eleanor. 1985. « L'évaluation de programmes aux États-Unis ». *Politiques et management public* 3(2) : 199-214.

Chelimsky, Eleanor. 1998. « The role of experience in formulating theories of evaluation practice ». *American Journal of Evaluation* 19(1) : 35-55.

Greene, Jennifer C. 1997. « Evaluation as advocacy ». *Evaluation Practice* 18(1) : 25-35. doi : https://doi.org/10.1177%2F109821409701800103

Madison, A., et V. Martinez. 1994. « Client participation in health planning and evaluation: An empowerment evaluation strategy ». Boston.

Scriven, Michael. 1973. « Goal-free evaluation ». in *School evaluation*, édité par E. R. House. Berkeley: McCutchan Publishing.

Scriven, Michael. 1991. *Evaluation Thesaurus*. 4e éd. Thousand Oaks: Sage Publications.

Shadish, William R., Thomas D. Cook et Laura C. Leviton. 1995. *Foundations of program evalution*. Thousand Oaks: Sage Publications.

4. La recherche transformationnelle : dimensions personnelles et sociétales

DONNA M. MERTENS

[Traduit de : Mertens, Donna M. 2017. « Transformative research: personal and societal ». *International Journal for Transformative Research*, 4(1) : 409-415 (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de De Gruyter.]

La recherche transformationnelle se concentre-t-elle sur la transformation de l'apprenant-e (ou plus généralement du participant ou de la participante) et du chercheur ou de la chercheuse, ou est-elle axée sur la transformation de la société?

Dans cet article, j'affirme que le fait de formuler cette question comme un choix entre deux options ne mènera pas à l'objectif de transformation souhaité. Au contraire, il serait plus utile de formuler la question sous l'angle de la complémentarité. Pour le dire autrement : Quelle est la nature de la recherche ayant un objectif transformationnel pour le participant ou la participante, le chercheur ou la chercheuse et la société? Et, si nous acceptons l'idée que l'objectif de transformation se situe à plusieurs niveaux, quelles sont les implications pour les méthodologies que nous utilisons pour mener cette recherche?

Lorsque je considère l'idée que la transformation nécessite l'imbrication des dimensions personnelle et sociétale, je pense à deux expériences que j'ai vécues personnellement. La première, c'est lorsque ma famille a déménagé de l'État de Washington à l'État du Kentucky au début des

années 1960, alors que je venais d'entrer en classe de cinquième. Tant que je vivais dans l'État de Washington, je n'avais jamais vu de personne Noire, mais leur présence m'est apparue immédiatement à mon arrivée dans le Kentucky. Ce que j'ai remarqué, c'est que les personnes Noires ne vivaient pas dans mon quartier et ne fréquentaient pas mon école ou ma piscine. La plus grande concentration de personnes Noires que j'ai vue vivait dans le centre-ville, sans climatisation, dans l'humidité étouffante du Kentucky. J'ai demandé à mon institutrice pourquoi les personnes Noires n'allaient pas à mon école. Elle m'a tapoté la tête et m'a dit : « Chérie, ils préfèrent juste être avec leurs semblables. » Je ne connaissais pas le mot dissonance cognitive à l'époque, mais c'est ce que j'ai ressenti. Ce fut un moment de transformation pour moi, car j'ai senti que quelque chose n'allait pas dans cette [façon de dépeindre les choses]. Sans en être pleinement consciente, c'est à ce moment-là que j'ai décidé de consacrer ma vie à découvrir ce qui n'allait pas dans cette représentation et ce qui pourrait être fait pour éliminer les discriminations qui limitaient les chances des personnes Noires et des membres d'autres communautés marginalisées dans la vie. Il s'agissait d'une transformation personnelle qui a débouché sur un engagement en faveur du changement sociétal.

La deuxième expérience personnelle s'est produite bien des années plus tard, lors d'une conférence à Amsterdam. J'avais terminé une présentation sur la recherche transformationnelle axée sur la transformation de la société, suivie d'une session de questions-réponses. Dr Bagele Chilisa, alors professeure agrégée de l'Université du Botswana, m'a demandé si j'avais envisagé la transformation des chercheurs et chercheuses eux/elles-mêmes. Sa question m'a prise par surprise; nous avons convenu de nous rencontrer pour en parler autour d'un dîner le soir même. À cette époque, j'étais en train d'écrire *Transformative Research and Evaluation* [La recherche et l'évaluation transformationnelles] (Mertens, 2009) et notre conversation m'a amenée à revoir le plan du livre. Je me suis rendu compte que je devais ajouter un chapitre entre l'introduction et le cadrage philosophique de la recherche transformationnelle et le chapitre sur le développement du sujet de recherche. Ce chapitre supplémentaire

s'intitule : *Self, Partnerships, and Relationships* [Soi, partenariats et relations] et porte sur la façon dont les chercheurs et les chercheuses peuvent arriver à se comprendre eux/elles-mêmes dans le contexte de la recherche de manière à faciliter l'établissement de relations de confiance avec les membres de la communauté de recherche. Ainsi, l'imbrication de la transformation personnelle avec l'objectif de transformation sociétale était un élément essentiel pour mieux comprendre comment mener une recherche sur la discrimination et l'oppression.

Je présente ces deux expériences en partie comme une réponse à la discussion de Walton (2014) sur le paradigme transformationnel dans laquelle elle décrit mon travail et ceux d'autres méthodologues transformationnel-le-s comme suit :

> Ces auteurs et autrices voient dans la recherche transformationnelle un moyen de réaliser des changements au niveau communautaire et institutionnel. Toutefois, la transformation peut également se produire au niveau personnel; et en effet, on peut faire valoir que la transformation à tout niveau doit commencer par la transformation de l'individu (p.30).

Nous constatons l'accent mis sur la transformation de l'individu dans des études telles que celle menée par Pratt et Peat (2014) qui porte sur la transformation d'un étudiant et d'un superviseur de thèse ou dans celle de Farren, Crotty et Kilboy (2015) dans laquelle ils ont étudié la transformation des enseignants et enseignantes par l'utilisation des technologies de l'information et de la communication dans un cours de langue.

Cet accent mis sur la transformation au niveau individuel est pertinent et nécessaire. Toutefois, il n'aborde pas nécessairement les questions qui font partie intégrante de la transformation au niveau sociétal. C'est une question soulevée par Walton qui suggère un lien entre l'engagement dans la recherche qui peut mener à une transformation personnelle et des transformations politiques, sociales et culturelles plus larges.

Walton et moi partageons peut-être plus de points communs qu'il n'y paraît dans la citation que j'ai fournie de son article de 2014. Nous appelons toutes deux à un changement dans la compréhension de la façon d'encadrer et de mener des recherches qui mènent à un monde plus juste. Nous reconnaissons toutes les deux que les approches traditionnelles de la recherche n'ont pas

> produit un monde durable ou une économie mondiale stable et équitable où chacun-e est nourri-e, soigné-e et éduqué-e… Donc, si les réalisations qui sont la conséquence du progrès de la science peuvent être saluées par ceux et celles qui vivent dans le confort matériel, il est important de garder à l'esprit que les souffrances d'un nombre incalculable de personnes continuent et que les problèmes de santé mentale, d'exploitation, de toxicomanie et de pauvreté existent dans les pays riches. Il est urgent d'évaluer radicalement les méthodes de recherche que nous utilisons et d'en créer de nouvelles qui s'attaqueront, aux niveaux individuel et collectif, aux crises sociales, écologiques et économiques urgentes qui menacent notre existence humaine. (Walton, 2014 : 40-41)

Mon travail sur le paradigme transformationnel en tant que cadre philosophique pour la recherche repose sur le postulat que si nous voulons contribuer à des changements transformationnels, nous devons consciencieusement concevoir nos recherches pour y incorporer cet objectif. J'émets l'hypothèse que la probabilité d'un changement transformationnel augmente lorsque nous reconnaissons explicitement que c'est notre objectif et que nous incluons des mécanismes dans la recherche pour soutenir ce changement. Par conséquent, je propose le paradigme transformationnel comme cadre pour concevoir une recherche qui intègre à la fois la transformation personnelle et sociétale.

Expériences de recherche personnelles qui ont conduit au développement du paradigme transformationnel

Au début de ma carrière de chercheuse, j'ai coordonné les travaux de recherche du *College of Medicine* de l'Université du Kentucky. À ce poste, j'ai publié un article qui remettait en question l'utilisation de bonnes notes dans les matières scientifiques comme principal critère d'admission des étudiants et étudiantes au programme. J'ai suggéré que d'autres critères pouvaient être pris en compte, tels que la capacité à établir des liens avec les gens et la représentation de divers groupes raciaux/ethniques et de genre. J'ai quitté ce poste pour coordonner l'évaluation d'un projet portant sur des zones de grande pauvreté aux États-Unis. Je me suis efforcée de représenter fidèlement les préoccupations des habitant-e-s de ces régions et j'ai cherché des méthodes pour faire en sorte que leurs intérêts soient entendus et pris en compte. De là, je suis allée à l'Université d'État de l'Ohio pour appuyer les décisions politiques relatives à l'enseignement professionnel. J'ai pu mener des études sur les expériences vécues par les personnes en situation de handicap, des décrocheurs et décrocheuses du secondaire, des étudiant-e-s des zones rurales isolées et des centres-villes, des femmes au travail et des détenu-e-s. Tout au long de ces expériences, j'ai éprouvé un profond sentiment d'inconfort parce que la plupart de mes recherches ont été menées à distance, à l'aide de bases de données existantes ou d'instruments d'enquête. Je savais que j'avais besoin de trouver un poste qui me permettrait de travailler avec les populations marginalisées, plutôt que « sur » elles.

À cette fin, j'ai accepté un poste d'enseignant à l'Université Gallaudet, la seule université au monde ayant pour mission d'être au service de la communauté sourde. J'ai commencé à y travailler avec l'idée que je voulais trouver comment entrer dans cette communauté marginalisée de façon respectueuse et faire des recherches avec cette communauté. Bien sûr, j'ai dû apprendre la langue des signes américaine et la culture

sourde. Ce que je n'avais pas prévu, et ce dont je suis très reconnaissante, c'est l'apprentissage qui s'est produit en moi sur la façon de mener des recherches respectueusement avec des personnes sourdes. Ces expériences m'ont amené à développer le paradigme transformationnel comme moyen d'intégrer les aspects de culture [propre à chaque communauté] et d'aborder les questions de discrimination et d'oppression de manière à permettre une transformation personnelle et sociétale.

En me plongeant dans le peu de documentation disponible à l'époque (au début des années 1980), je me suis rendu compte qu'il y avait un caractère unique aux expériences des Sourd-e-s, mais qu'ils et elles partageaient aussi des caractéristiques avec d'autres communautés marginalisées. De plus, la communauté sourde elle-même était hétérogène et représentait un microcosme du monde, dans lequel se retrouvait toute une variété de privilèges que peuvent vivre les gens en raison de leurs caractéristiques. En d'autres termes, les sourd-e-s viennent de pays, de groupes raciaux/ethniques, de genres, d'identités sexuelles et de milieux économiques différents. Par conséquent, j'ai cherché un moyen de comprendre comment élaborer une recherche qui tiendrait compte de la gamme complète des caractéristiques qui fondent les discriminations et l'oppression partout dans le monde. Ainsi, le paradigme transformationnel est né des préoccupations exprimées par les membres des communautés marginalisées et leurs défenseurs et défenseuses, qui estimaient que la recherche ne représentait pas fidèlement leurs expériences et ne contribuait pas de manière adéquate à l'amélioration de leurs conditions de vie (Mertens, 2015; Mertens et Tarsilla, 2015; Mertens et Wilson, 2012). L'impulsion est venue des communautés marginalisées qui ont constaté qu'une grande partie de l'évaluation était faite « sur » elles, mais elles ont noté que « peu de choses ont changé dans la qualité de vie des personnes qui sont pauvres et/ou discriminées en raison de leur race/ethnicité, de leur handicap, de leur surdité, de leur sexe, de leur indigénéité et d'autres dimensions pertinentes de la diversité » (Cram et Mertens, 2015 : 94, cité dans Mertens, 2018 : 21).

Les hypothèses philosophiques et les implications méthodologiques du paradigme transformationnel

Le paradigme transformationnel agit comme un cadre métaphysique réunissant des éléments philosophiques associés au féminisme, à la théorie critique, aux théories autochtones et postcoloniales, ainsi qu'aux théories des droits des personnes en situation de handicap et sourdes.

> Il s'applique aux personnes qui sont victimes de discrimination et d'oppression pour quelque raison que ce soit, y compris (mais non exclusivement) la race ou l'ethnicité, le handicap, le statut d'immigrant, les conflits politiques, l'orientation sexuelle, la pauvreté, le genre, l'âge ou la multitude d'autres caractéristiques associées à un moindre accès à la justice sociale. En outre, le paradigme transformationnel est applicable à l'étude des structures de pouvoir qui perpétuent les inégalités sociales. (Mertens, 2009 : 4)

Quatre hypothèses philosophiques constituent les éléments essentiels du paradigme transformationnel :

- L'axiologie ou la nature de l'éthique et des valeurs
- L'ontologie ou la nature de la réalité
- L'épistémologie ou la nature de la connaissance et la relation entre les chercheurs et chercheuses et ceux et celles qui participent ou sont affecté-e-s par la recherche
- La méthodologie ou la nature de l'investigation systématique

Ces quatre éléments ont été identifiés par Guba et Lincoln (2005) comme les hypothèses fondamentales qui guident les chercheurs et chercheuses dans leur processus d'investigation. La section suivante met en évidence la signification de ces hypothèses dans un paradigme transformationnel et intègre les niveaux de transformation personnels et sociaux qui correspondent à chaque hypothèse.

Hypothèse axiologique transformationnelle

Selon l'hypothèse axiologique transformationnelle, pour être éthique, une recherche doit être conçue de manière à promouvoir la justice sociale et les droits humains. Le point de départ de la recherche éthique est de comprendre ce que signifie le fait d'être respectueux de la culture des communautés dans lesquelles nous travaillons, de s'attaquer consciemment aux inégalités, de reconnaître les forces et la résilience d'une communauté et d'assurer la réciprocité aux membres de la communauté.

Le concept de respect culturel fournit une plateforme pour examiner l'imbrication des aspects personnels et sociétaux de la transformation. Les chercheurs et chercheuses occupent une position de privilège parce que

> leur rôle confère habituellement le pouvoir social de définir la réalité et de porter des jugements éclairés sur les autres… Les chercheurs et chercheuses ont la responsabilité éthique d'évaluer de façon proactive et d'aborder les façons dont notre répertoire personnel de ressources perceptuelles et interprétatives peut ignorer, obscurcir ou déformer plus qu'éclairer. (Symonette, 2009 : 280)

Le privilège est une position déterminée par la société, de sorte que le chercheur ou la chercheuse doit être conscient-e des dimensions de la diversité qui sont utilisées à la fois comme fondement du privilège et de la marginalisation.

Afin d'entreprendre des recherches respectueuses de la culture, les chercheurs et chercheuses doivent aussi examiner de façon critique leurs propres valeurs, croyances et hypothèses afin d'aller au-delà des « lunettes culturelles » qu'ils et elles apportent avec eux et elles dans le contexte de la recherche. La conscience de soi est nécessaire, mais pas suffisante; les chercheurs et les chercheuses doivent également faire

des efforts pour savoir comment ils et elles sont perçu-e-s par les participant-e-s d'une étude. Symonette pose la question essentielle : « Comment les personnes avec lesquelles vous cherchez à communiquer et à collaborer vous perçoivent-elles? » (p.289) La perception des participant-e-s à l'égard du chercheur ou de la chercheuse est une pièce cruciale du puzzle et déterminera la qualité des relations qui seront développées, ainsi que les données qui seront recueillies.

Walton (2014) met en exergue l'importance pour les chercheurs et chercheuses de dépasser les « lunettes culturelles » du matérialisme scientifique qui a dominé en Occident afin d'être ouvert-e-s aux significations de l'éthique qui proviennent des traditions spirituelles, religieuses et autochtones. L'imbrication des concepts de respect culturel et de spiritualité m'est apparue évidente lorsque j'ai travaillé avec deux chercheurs/-euses autochtones pour identifier les trajectoires par lesquelles les chercheurs/-euses autochtones doivent passer pour devenir des professionnel-le-s dans leur domaine. Les défis qu'ils et elles avaient rencontrés ne résultaient pas d'un manque de désir de recherche, mais plutôt d'une frustration liée au fait que leurs croyances culturelles n'étaient pas reconnues ou acceptées comme valides par de nombreux chercheurs et de nombreuses chercheuses externes, comme l'indique cette citation :

> Les méthodes de recherche autochtones sont aussi anciennes que les collines et les vallées, les montagnes et les mers, les déserts et les lacs auxquels les peuples autochtones sont liés en tant que lieux d'appartenance. Ce n'est pas que les peuples autochtones sont contre la recherche… la « mauvaise réputation » de la recherche au sein des communautés autochtones n'est pas liée à la notion de recherche elle-même; c'est plutôt la façon dont cette recherche a été pratiquée, par qui et dans quel but qui a créé des sentiments négatifs. (Cram, Chilisa, et Mertens, 2013 : 11)

Chilisa (2012) décrit un exemple de valeurs spirituelles de la communauté autochtone africaine. L'Ubuntu invite les chercheurs et chercheuses à mener leurs études en ayant conscience des effets de la recherche sur tous les êtres vivants et non vivants – ceux qui nous ont précédés, ceux qui sont avec nous maintenant et ceux qui viendront à l'avenir. Avec ce principe éthique directeur, comment les chercheurs et les chercheuses modifieraient-ils et elles la façon dont ils et elles conçoivent et mènent leurs recherches? Qu'implique cet impératif éthique pour nos méthodes de recherche si nous voulons nous assurer que nous ne nous contentons pas d'aborder la transformation personnelle, mais que nous contribuons également à des actions visant à transformer la société?

Hypothèse ontologique transformationnelle

L'hypothèse ontologique transformationnelle est qu'il existe de multiples versions de ce que l'on croit être réel et que ces croyances sont générées par de multiples facteurs. Les versions de la réalité proviennent de différentes positions sociétales associées à plus ou moins de privilèges, tels que le genre, l'identité sexuelle, la race, l'origine ethnique, la religion, le statut économique, le handicap et la surdité (Mertens, 2015). Il y a des conséquences associées à l'acceptation d'une version de la réalité plutôt qu'une autre. L'histoire que j'ai utilisée pour ouvrir ce chapitre donne un exemple d'une version de la réalité qui a été déterminée par les personnes Blanches de la classe moyenne. Ils et elles expliquent la ségrégation par le fait que les personnes Noires préfèrent rester avec « des gens de leur propre sorte ». Lorsque les personnes Noires et les défenseurs et défenseuses de l'équité raciale sont interrogé-e-s sur les raisons de la ségrégation, ils et elles décrivent une société qui pratique une discrimination fondée sur la couleur de peau et le pays d'origine d'une personne. La conséquence de l'acceptation de l'une de ces versions de la réalité plutôt que l'autre devrait être claire pour le lecteur ou la lectrice. Afin de soutenir la transformation de la société, les chercheurs

et chercheuses doivent également s'engager dans une transformation personnelle de leur compréhension des origines des différentes versions de la réalité et des conséquences de l'acceptation d'une version de la réalité plutôt qu'une autre.

Dans les premiers temps de la recherche en sciences sociales, la réalité était définie en termes de ce qui pouvait être observé et mesuré, écartant ainsi toute référence à des caractéristiques personnelles dans la collecte de données. Walton (2014) nous appelle à être plus ouvert-e-s aux possibilités relatives à la nature de la réalité en considérant la réalité qui provient de la reconnaissance des sentiments intérieurs, des intentions, des sentiments relatifs à ce qui fait sens et à la spiritualité. Elle propose que les chercheurs et les chercheuses en sciences sociales prêtent attention aux travaux de physique quantique et à leur conceptualisation d'une unité sous-jacente à la réalité. « Une conséquence est que la réalité existe en fin de compte comme une unité dans laquelle tout est intrinsèquement interconnecté; et notre perception sensuelle de la « séparation » dans le monde extérieur est une illusion. » (Walton 2014 : 34) Cette description de la réalité s'aligne sur le concept africain d'Ubuntu décrit plus haut dans cet article et a des implications pour la connexion entre les niveaux de compréhension personnel et sociétal.

Cette exploration du sens de la réalité et de ses sources nous amène à considérer le sens de la transformation elle-même. Qu'est-ce qui est accepté comme la réalité de la transformation? Cette question a des réponses différentes en fonction de la personne à qui on la pose. Dans le contexte de l'apprentissage transformationnel, Smith (2016) a décrit la transformation dans la salle de classe en décrivant la façon dont les conférenciers et conférencières ont transformé leur approche de l'enseignement par des utilisations créatives de la technologie. Jones (2015) a décrit la transformation dans la vie de jeunes privés de leurs droits par le passage d'une vie de victimes de négligence et de maltraitance, à une vie où ils et elles sont capables de s'épanouir en tant que jeunes personnes confiantes, ayant un sens positif de l'identité et de l'estime de soi. Hammond (2016) a décrit la transformation des

enseignant-e-s par la tenue de blogues mettant l'accent sur la réflexivité critique. Ces transformations se concentrent sur le niveau individuel tout en ayant des implications sociales plus larges.

Lorsque les peuples autochtones sont interrogés sur la transformation, ils décrivent la nécessité de la décolonisation tant au travers des méthodes de recherche que par la restitution de leurs terres, de leurs ressources et des libertés qui leur ont été retirées (Cram et Mertens, 2015). Il s'agit d'une transformation clairement axée sur le niveau sociétal; mais les peuples autochtones soulignent qu'une telle transformation doit passer par l'établissement de relations entre eux et avec les non-Autochtones. Lorsqu'on a demandé aux personnes présentant une déficience intellectuelle quelles étaient leurs priorités en matière de transformation, elles ont répondu qu'elles souhaitaient vivre dans un monde où elles pouvaient mener une vie « ordinaire » (National Health Committee, 2003). Cette définition de la transformation a des implications au niveau sociétal et personnel. La transformation doit porter sur les attitudes et les obstacles sociétaux qui limitent les chances des personnes handicapées dans la vie, mais aussi inclure la transformation personnelle des personnes au pouvoir et des personnes en situation de handicap.

Hypothèse épistémologique transformationnelle

L'hypothèse épistémologique transformationnelle se concentre sur la signification de la connaissance telle qu'elle est perçue à travers de multiples « lunettes culturelles » et sur l'importance des inégalités de pouvoir dans la reconnaissance de ce qui est considéré comme une connaissance légitime (Mertens, 2015). Cela signifie que les chercheurs et les chercheuses doivent être conscient-e-s de leur propre pouvoir et de leurs « lunettes culturelles », et de la façon dont ils et elles influencent leurs relations avec les participant-e-s de la recherche. Comme me l'ont appris les chercheurs et chercheuses autochtones et les membres

d'autres communautés marginalisées, tout est dans la relation. Sur le plan personnel, les chercheurs et chercheuses doivent transformer leur façon de pénétrer respectueusement dans les communautés afin d'établir des relations qui reconnaissent les connaissances que les membres de la communauté apportent. Par exemple, en tant que personne non sourde qui effectue des recherches auprès de la communauté sourde depuis plus de 30 ans, j'ai dû modifier ma perception de moi-même en tant qu'experte dans les contextes de recherche pour reconnaître qu'en matière de surdité, je ne suis pas l'experte. Les personnes qui ont vécu une expérience de la surdité sont les expertes à cet égard et cette connaissance doit être reconnue et valorisée. Les chercheurs et les chercheuses ont la responsabilité de concevoir des stratégies qui permettent à ceux et celles qui détiennent le pouvoir traditionnel et à ceux et celles qui ont été exclus du pouvoir d'être impliqué-e-s de manière respectueuse. Cela exige des chercheurs et chercheuses qu'ils et elles transforment leur rôle pour soutenir l'expertise qui existe dans les communautés dans lesquelles elles travaillent de manière significative.

Cette transformation pourrait aussi impliquer une prise de conscience et une appréciation croissante des types de connaissances transpersonnelles qu'Anderson et Braud (2011) associent aux transformations obtenues, mais qui ne sont généralement pas inclus dans la recherche en sciences sociales. Il s'agit notamment (1) des connaissances intuitives que nous avons, sans attendre la prise de conscience de l'esprit rationnel; et (2) des connaissances profondes qui incorporent les connaissances relatives à la discipline de recherche aux connaissances tacites, intuitives, corporelles et basées sur les sentiments pour soutenir la croissance psycho-spirituelle. Cela est conforme à l'hypothèse du paradigme transformationnel sur le savoir, en ce sens que ces types de connaissances sont valorisés par différents groupes culturels. Par exemple, les chercheur·es autochtones valorisent les connaissances qui sont enracinées dans une spiritualité qui se manifeste par une connexion avec tout ce qui est venu avant, tout ce qui est ici

maintenant et tout ce qui sera. Ces connaissances peuvent être transmises aux membres de la communauté sous diverses formes, même sous forme de rêves (Cram *et al.*, 2013).

Hypothèse méthodologique transformationnelle

L'hypothèse méthodologique transformationnelle ne dicte aucune approche méthodologique spécifique. Elle s'aligne plutôt sur les hypothèses transformationnelles discutées précédemment, en ce sens que les voix des personnes marginalisées dans la société doivent être intégrées de manière significative dans la planification et la mise en œuvre de la recherche. Cela signifie qu'une analyse des relations de pouvoir doit être effectuée dans le cadre du processus d'orientation de la recherche, ainsi que tout au long du processus de recherche. Il est très important d'intégrer délibérément dans la conception de la recherche des moyens pour permettre une transformation personnelle et sociétale. Cela n'est pas laissé au hasard.

À cette fin, les chercheurs et chercheuses transformationnel-le-s adoptent souvent une approche itérative en méthodes mixtes, en utilisant les premières étapes de la recherche pour identifier qui doit être inclus et comment ils et elles peuvent l'être (Mertens, 2018). Cela implique également un processus de transformation au niveau individuel qui favorise l'établissement de relations de confiance et la collaboration avec les membres de communautés marginalisées et en situation de pouvoir afin de comprendre les complexités culturelles à l'œuvre et leurs implications pour la transformation. La phase d'établissement de relations peut être suivie d'une phase d'analyse contextuelle au cours de laquelle les données et la documentation existantes peuvent être examinées. Elle peut aussi inclure des stratégies faisant appel à des processus collectifs pour mettre en lumière les types de connaissances tacites et profondes qui constituent la base de la transformation. Les informations recueillies

au cours de ces phases sont utilisées pour élaborer une intervention susceptible de transformer les individus et la société. Cette intervention fait généralement l'objet d'un essai pilote auprès d'un petit groupe afin de pouvoir l'adapter si nécessaire. Les recherches portant sur la mise en œuvre peuvent utiliser divers modèles, allant de l'étude de cas à la recherche sur l'action participative en passant par des expérimentations par assignation aléatoire. Elles doivent être menées avec une pleine compréhension de la signification de l'éthique dans un contexte de recherche transformationnelle. Pendant la phase de collecte, des données sont recueillies sur les processus et les résultats de l'intervention. Lorsque les données finales sont recueillies sur les effets de l'intervention, elles sont rapportées à la communauté en faisant appel à diverses stratégies d'interprétation et d'utilisation des résultats. Cette utilisation peut consister dans des changements transformationnels dans une école ou une classe donnée ou dans un changement de politique susceptible d'affecter un groupe plus large.

L'hypothèse méthodologique transformationnelle s'aligne sur la recommandation de Walton (2014) de recueillir des données auprès de multiples sources de diverses manières qui honorent les connaissances intuitives et profondes, nécessaires à la transformation. Walton suggère que les méthodes transpersonnelles sont intéressantes pour former les chercheurs et les chercheuses et les participant-e-s à reconnaître les connaissances intuitives. Cela peut inclure la mise en place d'un dialogue imaginaire lors de l'élaboration du sujet de recherche, l'utilisation d'une stratégie dans laquelle un vaste examen de la littérature permet de remettre en question les valeurs et les hypothèses personnelles, la combinaison de méthodes de recherche intuitives avec des méthodes conventionnelles quantitatives, qualitatives et mixtes, et l'intégration des résultats de toutes les collectes de données avec la littérature et leur partage de manière significative avec divers publics. « La perception intuitive peut aider à atteindre des formes de compréhension plus riches

lorsqu'elle est utilisée pour compléter des processus tels que le raisonnement analytique et l'information obtenue à travers les cinq sens conventionnels » (p.37).

Conclusions

Le paradigme transformationnel fournit un cadre philosophique pour la conception de recherches susceptibles d'apporter des changements aux niveaux individuel et sociétal. En ce qui me concerne, ce cadre m'incite à dialoguer différemment avec les participant-e-s aux études menées, à poser différents types de questions de recherche et à concevoir des études visant à soutenir des changements qui remettent en question un statu quo oppressif. L'inclusion de connaissances fondées sur l'intuition et les rêves n'élimine pas l'importance des connaissances issues de méthodes plus traditionnelles de collecte de données. Elle permet de tenir compte des différences de compréhension culturelle de ce qu'est le savoir et offre la possibilité de parvenir à une compréhension plus riche de la signification des expériences et des changements.

Je suis d'accord avec Walton (2014) et Anderson et Braud (2011) sur la nécessité d'une conceptualisation différente de la méthodologie de recherche afin de tenir compte de la diversité culturelle et des différents modes de connaissance. J'ajoute à leur réflexion la nécessité de concevoir des études qui abordent explicitement les questions de discrimination et d'oppression. Le changement individuel est un objectif souhaitable; toutefois, les personnes qui subissent une discrimination systémique constatent que leurs chances dans la vie sont limitées par un système oppressif. Il est donc nécessaire de s'attaquer à la fois à l'individu et à la société dans le cadre de la recherche transformationnelle.

Je suis également d'accord avec Walton (2014) et Anderson et Braud (2011) sur le fait que l'utilisation proposée de stratégies transformationnelles inclusives dans la recherche ne nie pas l'importance de ce que l'on sait

des bonnes pratiques de recherche. Les stratégies transformationnelles peuvent compléter et améliorer les approches de recherche traditionnelles. Les chercheurs et les chercheuses transformationnel-le-s soutiennent l'utilisation de multiples méthodes pour la réalisation d'études, ainsi que l'élaboration d'approches interdisciplinaires pour résoudre des problèmes difficiles. Je pense que l'intégration des concepts de changement personnel et sociétal sera bénéfique pour le monde. Je termine par cette citation de Walton (2014 : 36) qui saisit l'essence de cet argument :

On ne cesse de mettre l'accent sur la nécessité d'un pluralisme méthodologique, où les chercheurs et les chercheuses d'un éventail de disciplines, y compris les sciences sociales, les sciences naturelles, les sciences humaines et les arts, peuvent adopter des approches individuelles et collaboratives pour générer des connaissances qui aborderont des questions d'intérêt mondial.

Bibliographie

Anderson, Rosemarie, et William Braud. 2011. *Transforming Self and Others through Research: Transpersonal Research Methods and Skills for the Human Sciences and Humanitie*. New York: SUNY Press.

Chilisa, Bagele. 2012. *Indigenous methodologies*. Thousand Oaks: Sage Publications.

Cram, Fiona, Bagele Chilisa, et Donna M. Mertens. 2013. « The journey begins ». in *Indigenous pathways into social research*, édité par F. Cram, B. Chilisa et D. M. Mertens. Walnut Hills: Left Coast Press, p. 11-40.

Cram, Fiona, et Donna M. Mertens. 2015. « Transformative and Indigenous frameworks for multimethod and mixed methods research ». in *The Oxford handbook of multimethods and mixed methods research inquiry*, édité par S. N. Hesse-Biber et R. B. Johnson. New York: Oxford University Press, p. 91-110.

Farren, Margaret A., Yvonne Crotty et Laura Kilboy. 2015. « Transformative potential of action research and ICT in the second language (L2) classroom ». *International Journal of Transformative Research* 2(2) : 49-59.

Hammond, Michael. 2016. « How ideas of transformative learning can influence academic blogging ». *International Journal of Transformative Research* 3(1) : 33-40.

Jones, Jocelyn. 2015. « Professional engagement in child protection: promoting reflective practice and deeper connection with the lived reality for children ». *International Journal of Transformative Research* 2(2) : 30-38.

Mertens, Donna M. 2009. *Transformative research and evaluation*. New York: Guilford Press.

Mertens, Donna M. 2015. *Research methods in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. 4e éd. Thousand Oaks: Sage Publications.

Mertens, Donna M. 2018. *Mixed methods design in evaluation*. Thousand Oaks: Sage Publications.

Mertens, Donna M., et Michele Tarsilla. 2015. « Mixed methods evaluation ». in *The Oxford handbook of multimethod and mixed methods research inquiry*, édité par S. Hesse- Biber et R. B. Johnson. Oxford: Oxford University Press, p. 426-46.

Mertens, Donna M., et Amy T. Wilson. 2012. *Program Evaluation Theory and Practice: A Comprehensive Guide*. New York: Guilford Press.

National Health Committee. 2003. « To have an “ordinary” life, Kia whai oranga “noa.” »

Pratt, Dee, et Beth Peat. 2014. « Vanishing point – or meeting in the middle? Student/supervisor transformation in a self-study thesis ». *International Journal of Transformative Research* 1(1) : 1-24.

Smith, David. 2016. « An intuitive approach to learning delivery in higher education ». *International Journal of Transformative Research* 3(2) : 8-14.

Symonette, Hazel. 2009. « Cultivating self as responsive instrument ». in *Handbook of Social Research Ethics*, édité par D. M. Mertens et P. E. Ginsberg. Thousand Oaks: Sage Publications, p. 279-94.

Walton, Joan. 2014. « What can the ‘transpersonal’ contribute to transformative research? » *International Journal of Transformative Research* 1(1) : 25-44.

5. L'évaluation contribue-t-elle au bien commun?

SANDRA MATHISON

[Traduit de Mathison, Sandra. 2018. « Does Evaluation Contribute to the Public Good? », *Evaluation*, 24(1) : 113-119 (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Introduction

Si nous voulons que l'évaluation apporte une contribution positive au monde social, physique et environnemental, il faut commencer par analyser notre théorie et notre pratique d'un point de vue sociologique. Les idéologies sociopolitiques dominantes façonnent la conceptualisation de l'évaluation, les méthodes et les modèles utilisés, comment et par qui elle est financée, ainsi que son efficacité à promouvoir un changement social positif. La plupart des évaluations se tiennent dans un contexte micro, un héritage de la pratique évaluative qui se considère au service d'autres disciplines, des décideurs, des politiques, des bailleurs ou des bénéficiaires des programmes. La pratique de l'évaluation est locale (même lorsque le contexte est géographiquement vaste) et surtout attentive à répondre à des préoccupations portant sur l'efficacité des programmes évalués.

Les évaluateurs, tout en continuant à travailler dans le cadre des programmes, devraient examiner les cadres eux-mêmes. Nous devons nous demander comment ces cadres tiennent pour acquis et soutiennent une certaine façon de définir les problèmes, les solutions apportées et

les indicateurs de succès. Ces cadres s'inscrivent dans des idéologies qui structurent les relations humaines et les pratiques sociales au-delà de l'évaluation, mais également en son sein.

Je retracerai l'évolution de la théorie et de la pratique de l'évaluation sous l'influence d'idéologies mondiales, du progressisme initial (caractérisé par le financement public d'une grande partie des évaluations de programme) à un néolibéralisme peut-être en déclin (caractérisé par un financement accru par les organisations philanthropiques, les ONG et les entrepreneurs) jusqu'à un populisme en plein essor, et je réfléchirai à la contribution de l'évaluation au bien commun à travers ces changements.

Bien qu'il s'agisse peut-être d'une remarque déplaisante, nous devons nous demander si l'évaluation, telle qu'encadrée par ces idéologies sociopolitiques dominantes, contribue au bien commun, si elle contribue à un changement positif. De l'avis général, le travail des évaluateurs et des évaluatrices ne contribue pas suffisamment à la réduction de la pauvreté, aux droits humains, à l'accès à la nourriture, à l'eau, à l'éducation et aux soins de santé. Nous devons également nous demander si la pratique de l'évaluation classique ne risque pas d'entraver et de freiner le changement social. Je conclurai par quelques pistes de réflexion sur ce que nous (évaluateurs/-trices, bailleurs/-euses de fonds et utilisateurs/-trices de l'évaluation de programme) pourrions faire pour apporter une contribution positive au bien commun par le biais de l'évaluation.

Je fais des évaluations, essentiellement dans le domaine de l'éducation, depuis plus de 40 ans. J'aimerais pouvoir dire que l'évaluation a rempli sa promesse d'amélioration [de l'action publique]. Que la pratique comme la théorie en évaluation ont amélioré les choses. Mais au lieu de cela, il se trouve que je suis pessimiste, peut-être même cynique quant à la contribution de l'évaluation au bien commun, que je définis comme le bien-être de tous, dans le monde entier, et qui se manifeste par des choses telles que la sécurité alimentaire, l'accès aux soins, à l'éducation, à l'eau potable et au logement. Même si je ne présume pas que l'évaluation ainsi que les évaluateurs et les évaluatrices seraient seul-e-s responsables

du bien commun, en tant qu'évaluateurs et évaluatrices nous suggérons que le travail que nous faisons améliorera les choses, mènera à de meilleurs résultats, résoudra des problèmes sociaux et environnementaux. Nous devons donc prendre nos responsabilités en conséquence. Même si je suis assez pessimiste, je terminerai par quelques idées auxquelles nous pourrions réfléchir si nous croyons toujours à la promesse dont l'évaluation est porteuse.

Examinons un certain nombre de conditions actuelles dans le monde :

- Le monde a une population de 7,5 milliards de personnes
- La moitié de la population mondiale vit dans la pauvreté
- 22 000 enfants meurent chaque jour parce qu'ils vivent dans la pauvreté
- 805 millions de personnes n'ont pas de quoi se nourrir
- 5 millions d'enfants meurent chaque année avant l'âge de 5 ans
- 165 millions d'enfants de moins de 5 ans souffrent d'un retard de croissance dû à la malnutrition
- 750 millions de personnes ne disposent pas d'un accès adéquat à l'eau potable
- 2 300 personnes meurent chaque jour de diarrhée
- 214 millions de femmes n'ont pas accès à la planification familiale
- 1 enfant sur 7 à New York est sans abri
- Cette année, en Colombie-Britannique, 4 personnes mourront chaque jour d'une overdose d'opiacés
- Chaque jour, 34 500 personnes fuient leur foyer pour éviter la violence
- 1 personne sur 113 sur la Terre a été chassée de chez elle par les conflits, la violence, ou des violations des droits de l'homme
- Deux tiers des personnes analphabètes dans le monde sont de sexe féminin.

Je vais mettre en avant trois raisons pour lesquelles je crois que l'évaluation n'a pas contribué et ne contribue pas assez au bien commun.

Premièrement, la théorie et la pratique évaluative (comme de nombreuses autres pratiques sociales) reflètent les valeurs, les croyances et les préférences de l'époque. À ce titre, l'évaluation est contrainte par les idéologies sociales et politiques dominantes.

Deuxièmement, l'évaluation manque fondamentalement d'indépendance : c'est une prestation de service fournie à ceux qui ont le pouvoir et l'argent, et dans cette relation, elle devient une pratique limitée dans sa capacité à contribuer au bien commun.

Troisièmement, l'évaluation est fondamentalement une pratique conservatrice, s'inscrivant au sein de domaines [disciplinaires] établis par d'autres et, le plus souvent, maintenant le *statu quo*.

L'influence des idéologies sociopolitiques

La façon que nous avons de porter des jugements sur la valeur et la qualité des programmes, des politiques, des interventions et des réformes est en fonction d'idéologies sociopolitiques. Les idéologies sociopolitiques dominantes façonnent la façon dont l'évaluation est menée, conceptualisée, les méthodes et modèles utilisés, comment et par qui elle est financée, et sa capacité à promouvoir un changement social positif. Toute évaluation nécessite de définir les qualités souhaitables de ce qui est évalué. Et ces qualités sont socialement construites; par conséquent, les approches dominantes de l'évaluation reflètent l'esprit sociopolitique de notre époque.

Dans notre histoire récente, deux idéologies sociopolitiques dominantes ont façonné la manière dont une évaluation est conceptualisée et réalisée : le progressisme (*progressivism*) ou social-démocratie, et le néolibéralisme. Et, nous sommes potentiellement au bord d'une autre idéologie dominante : le populisme.

J'utilise le terme progressisme pour cette première phase même si, en réalité, l'ère sociopolitique progressiste a une histoire beaucoup plus longue. J'aurais aussi pu utiliser le terme de social-démocratie. L'idée principale à laquelle je me rattache est le plaidoyer en faveur de la réforme sociale, en particulier la réduction des inégalités de revenus, l'élargissement des libertés et des droits, et des expressions diverses de l'action collective et humanitaire l'idée que la condition humaine s'améliorerait grâce à la science, à la technologie et à l'organisation sociale.

À ses débuts, la pratique de l'évaluation portait la marque du progressisme. Notre travail était souvent mené sur fonds publics et se définissait comme un bien commun, au service de l'intérêt général. L'évaluation reflétait des valeurs progressistes… y compris l'efficacité, la justice sociale et la démocratie. La fin des années 1960 et le début des années 1980 ont été pour l'évaluation l'équivalent de la Ruée vers l'Or. À cette époque, les approches évaluatives se sont multipliées et un travail intellectuel passionnant a été effectué dans un certain nombre de disciplines. Les évaluateurs et les évaluatrices ont emprunté à d'autres disciplines (comme le journalisme, la jurisprudence, l'art) pour explorer comment nous pouvions et devrions émettre des jugements sur la valeur [des interventions évaluées]; les évaluateurs et les évaluatrices ont exploré le potentiel de l'évaluation pour contribuer à la démocratie, à la justice sociale et à l'équité; les évaluateurs et les évaluatrices étaient convaincu-e-s que leur métier les aiderait à rendre le monde meilleur.

Cela fait maintenant plusieurs décennies que l'évaluation peine sous l'ère néolibérale, et dans l'état actuel du néolibéralisme, l'évaluation reflète de plus en plus ses valeurs, y compris la marchandisation, la concurrence et la privatisation. Le néolibéralisme est une idéologie sociopolitique mondiale sans pareil… Il rejette les débats politiques partisans traditionnels et ignore les frontières nationales.

Notre époque est celle où le capitalisme éclipse la démocratie, et où l'interdépendance entre capital et gouvernement devient patente. Un principe fondamental du néolibéralisme est le rôle central que les gouvernements des États jouent dans la facilitation et la promotion des intérêts des élites économiques.

Et l'évaluation est une pratique qui, d'une part, est un outil pour rationaliser et normaliser l'action de l'État et les valeurs néolibérales. C'est manifeste, par exemple, dans le Nouveau management public… la gestion de la performance… la mesure de l'impact. Nous le voyons dans la montée du philanthrocapitalisme, qui apporte ses stratégies agressives, ses mesures de performance, et met l'accent sur l'efficacité [à court terme] dans le secteur à but non lucratif. Nous voyons cela dans la montée de l'investissement social, une confusion délibérée entre « faire le bien », et réaliser un profit.

D'autre part, l'évaluation est devenue l'outil de l'État qui surveille et évalue constamment les politiques publiques, la conduite des organisations, des agences et des individus, devenant même le juge ultime. Et de fait, l'évaluation est le fournisseur privilégié de connaissances pour l'individu, l'organisation ou l'État rationnels qui cherchent des données impartiales, comparables et arrivant juste à temps pour faire des choix rationnels parmi des solutions concurrentes. Enfin, la prise de décision basée sur des faits, plutôt que sur l'habitude, la préférence, la communauté ou la magie.

Un bon exemple de l'État comme évaluateur est le Centre d'information Clearinghouse sur « ce qui marche » (*What works Clearinghouse*) créé par le ministère de l'Éducation américain pour dire aux enseignants et aux parents quels programmes, produits, pratiques et politiques éducatifs marchent et lesquels ne marchent pas. Ce qui marche est directement connecté à des conceptions étroites de la façon dont on sait ce qui fonctionne (dans ce cas, les seules preuves qui comptent sont celles qui sont issues d'études contrefactuelles, avec une préférence pour l'approche expérimentale).

Le néolibéralisme crée donc à la fois un État évalué et un État évaluateur.

Malgré des réflexions sincères, bien argumentées et convaincantes défendant une vision de l'évaluation de programmes comme démocratisante, émancipatrice, inclusive et transformatrice, et/ou participative, la pratique de l'évaluation reflète souvent (peut-être généralement) l'accent mis sur l'efficience, l'efficacité et des résultats mesurables à court terme typiquement reflétés dans une logique d'intervention ou une théorie du changement. L'évaluation adopte la « langue du marché » et mesure le succès des programmes en termes de profit; les programmes sont en compétition; les résultats sont conceptualisés en termes économiques; et si c'est un échec, ce sont les individus qui sont à blâmer.

Les évaluateurs et les évaluatrices parlent d'un bon programme comme présentant un bon retour sur investissement ou un bon rapport qualité-prix, qui réduit les coûts (que l'on parle de santé, de logement ou de la sécurité sociale) ou permet de faire des économies, et qui est le meilleur en son genre. Et lorsque les programmes ne fonctionnent pas ou échouent, c'est que des prestataires de services n'ont pas su mettre en œuvre fidèlement les consignes qui leur sont données, ou que les bénéficiaires n'étaient pas suffisamment volontaires ou motivés. Même lorsque l'évaluation est formative, le rôle de l'évaluation reste souvent de mettre en avant ce qui doit être fait pour qu'une intervention soit généralisée ou reproduite ou pour qu'elle touche plus de gens.

Bien qu'une grande partie du monde soit toujours sous l'emprise du néolibéralisme, une idéologie émerge, même si nous ne savons pas bien ce que seront sa portée et sa puissance dans l'avenir. Ce que les grands médias appellent le populisme de droite, mais qu'il conviendrait plus précisément d'appeler l'autoritarisme ou même le néofascisme, traverse les sociétés démocratiques, notamment la Grande-Bretagne, la France, la Turquie et les États-Unis.

Tout populisme, qu'il soit de droite ou de gauche, a en commun l'idée de rassembler des gens qui se sentent mal ou sous-représentés, autour d'une vision du « nous » contre « eux ». Le populisme peut être une réponse au néo-libéralisme et en même temps un appel au centralisme et l'espoir que le compromis, le développement d'un consensus au centre, l'emportera.

Je ne sais pas ce à quoi l'évaluation ressemblera dans un monde où le populisme de droite serait dominant. Mais ce que j'ai pu constater au cours de mes 40 années d'évaluation, c'est que nous pouvons être assurés que l'évaluation sera modelée au sein de cette idéologie sociopolitique d'une manière qui serve ses principes et ses valeurs centrales.

L'évaluation manque fondamentalement d'indépendance

L'évaluation elle-même est une marchandise, une prestation de service, et ceci est particulièrement exacerbé dans un cadre capitaliste néolibéral. L'évaluation est un service acheté et vendu et bien que de nombreux évaluateurs et de nombreuses évaluatrices encadrent leur pratique dans les grands principes [déontologiques] de nos organisations professionnelles, l'évaluation et les évaluateurs/-trices sont néanmoins sensibles [aux attentes de] ceux et celles qui paient pour leurs services. Je crois qu'il est difficile pour la plupart des évaluateurs et des évaluatrices qui pratiquent de réellement imaginer que celui ou celle qui a commandé l'évaluation n'a pas le dernier mot sur les questions d'évaluation et les effets à évaluer.

Ceux et celles qui ont l'argent dominent la définition de ce qui est important, de ce qui compte comme succès et de la façon dont cela est démontré. Le mode de fonctionnement dans une grande partie de la pratique évaluative est descendant, ce que Michael Scriven a appelé l'« idéologie managériale ». En élargissant la focale, les gestionnaires des

programmes / projets servent leurs bailleurs et bailleuses de fonds et les évaluateurs/-trices servent les gestionnaires de programme / projet et/ ou leurs bailleurs/-euses de fonds.

Il est à noter que l'un des sujets qui a fait l'objet du plus grand nombre de recherches, dont on a le plus discuté et dont on s'est le plus inquiété dans l'évaluation c'est son UTILISATION, ou plus précisément son absence d'utilisation. Nous nous sommes interrogés sans relâche pendant des dizaines d'années sur les raisons pour lesquelles nos évaluations ne sont pas utilisées, comment nous pouvons amener les gens à les utiliser, comment nous pouvons faire une évaluation pour que les décideurs et les décideuses voient l'utilité de notre travail. Carol Weiss nous a d'abord libérés de l'idée que les résultats de l'évaluation mèneraient directement, de façon instrumentale, aux décisions et aux changements. Mais, ce faisant, elle nous a aussi laissé la désagréable sensation que notre travail était une petite partie d'un grand tout, ou pire encore, un pion dans un jeu joué selon des règles complètement différentes [des nôtres]. Récemment, nous avons même créé un nouveau type d'utilisation, l'utilisation processuelle, pour nous convaincre, nous et les autres, que ce que nous faisons est vraiment important.

Toute profession qui dépense tant d'énergie, tant de temps, tant de ressources sur les raisons pour lesquelles les gens ne s'intéressent pas à ce que nous faisons et disons est par nature une profession qui ne fait sans doute pas autant de bien qu'on pourrait le croire.

Parce que nous sommes pris au piège de cette idéologie managériale, nous sommes fier-e-s de fournir des données probantes sur ce qui marche et ce qui ne marche pas, que nous défendons souvent face à des soutiens irrationnels ou du moins idéologiques envers [certaines] politiques et programmes.

On entend l'expression « dire la vérité au pouvoir », utilisée pour exprimer ce que nous considérons comme notre contribution à faire ce qui est juste, à contribuer au bien commun. Un acte courageux, mais le plus

souvent futile. « Dire la vérité au pouvoir » est un cliché, utilisé souvent par les gauchistes et les progressistes, et qui laisse de côté la possibilité que les puissants et les puissantes sachent déjà la vérité et choisissent de l'ignorer (ou de la modifier) pour soutenir leurs intérêts et leurs idéologies déjà bien développées. Malheureusement, notre dépendance à l'égard des autres dans notre rôle de prestataire de service et notre longue expérience visant à traduire ce que nous avons appris dans les compréhensions, les préférences et les présupposés des autres font que dire la vérité au pouvoir n'est vraiment pas la bonne approche, ou du moins une approche pas très efficace.

L'évaluation est conservatrice

La plupart des évaluateurs pensent dans un microcontexte : c'est là l'héritage d'une pratique évaluative [...] au service des décideurs/-euses, des politiques, des bailleurs/-euses et des bénéficiaires. La pratique de l'évaluation est locale (même lorsque le contexte est géographiquement vaste) et répond principalement à des préoccupations particulières portant sur l'efficacité des programmes évalués. L'apport de l'évaluation pour réduire la pauvreté et ses conséquences sur la population serait typiquement d'identifier des stratégies efficaces, ou d'améliorer les stratégies existantes de réduction de la pauvreté. Bien qu'il soit tout à fait approprié que les praticiens et les praticiennes de l'évaluation travaillent de cette manière, cela détourne notre attention des grandes questions sur pourquoi cette intervention, pourquoi cette stratégie, pourquoi ces personnes et non pas d'autres.

Prenez cet exemple simple : la sécurité alimentaire, un concept fondamental en matière de réduction de la pauvreté, une contribution essentielle au bien commun. Les programmes visant à accroître la sécurité alimentaire sont créés par des organismes (comme *Save the Children*) ou les gouvernements (comme *USAID* [L'agence bilatérale d'aide

au développement américaine]) qui à leur tour sont les architectes des critères de succès de leurs programmes. *USAID* déclare que les évaluations devraient mettre l'accent sur la gestion axée sur la performance pour « renforcer l'impact de ces programmes sur le bien-être de leurs bénéficiaires ». Bien qu'il existe de nombreux types de programmes de sécurité alimentaire, l'un d'entre eux est le programme « Nourriture contre travail » (*food for work*), une stratégie dans laquelle le paiement de l'aide alimentaire est échangé contre du travail dans les programmes de travaux publics conçus pour construire et entretenir l'infrastructure locale (comme les routes, les puits, les latrines ou les écoles).

Ce type de programme (une approche d'échange de marchandises contre un besoin humain fondamental, dans ce cas la nourriture) privilégie la croissance économique locale comme résultat principal, et la sécurité alimentaire est un moyen vers cette fin. Les bénéficiaires sont à la fois des capitalistes et des personnes qui ont besoin de nourriture. Une réponse à la sécurité alimentaire qui mettrait l'accent sur la valeur d'usage (plutôt que sur la valeur d'échange) [de la nourriture] conduirait à différentes stratégies et indicateurs de succès; par exemple, la valeur d'usage de la nourriture contribue à maintenir les liens familiaux, l'amour, l'esthétique, le bonheur, le bon voisinage et le développement communautaire.

Mais le travail de l'évaluateur ou de l'évaluatrice est pris dans ce genre de système fermé et il faudrait un effort herculéen pour mettre plutôt l'accent de l'évaluation sur les hypothèses sous-jacentes d'un programme. Par conséquent, la pratique de l'évaluation est une réaction aux idées des autres et n'a qu'une portée limitée pour contester l'idée de la politique ou du programme.

Ajoutez à cela la probabilité que nous fassions des évaluations de programmes dont les hypothèses et les intentions sont en accord avec nos propres systèmes de valeurs, et les chances de remettre sérieusement en cause les intentions sous-jacentes des programmes sont encore diminuées.

Notre travail a donc tendance à conserver ce qui est déjà là.

Quelques réflexions pour aller de l'avant

Nous avons besoin d'évaluations réellement indépendantes

L'évaluation externe est souvent considérée comme plus indépendante et susceptible de fournir un diagnostic sans préjugés sur les programmes, les interventions et les stratégies. La question de savoir si cela est vrai a fait l'objet de nombreux débats, mais a minima, les évaluations payées par quelqu'un ayant un intérêt dans le programme seront influencées par ces intérêts. Je suggère que des évaluations plus indépendantes, effectuées sans intérêt monétaire dans le programme (mais avec des intérêts intellectuels ou moraux) puissent fournir des informations sur la valeur et les conséquences des programmes et des interventions. Et qu'ainsi, le financement indépendant de ces évaluations permette aux évaluateurs d'intervenir en dehors des cadres du néo-libéralisme, de regarder les résultats à long et à court terme, et d'enquêter sur les résultats non planifiés et non anticipés. J'ai utilisé précédemment l'exemple des programmes de sécurité alimentaire conçus comme un échange de produits et montré que la façon dont à la fois le problème (la faim, la famine) et la solution (la nourriture en échange du travail) étaient envisagés n'était pas a-idéologiques, mais basés sur les principes du capitalisme. Une évaluation indépendante qui interrogerait les fondements de ces programmes serait une contribution substantielle à la remise en question d'hypothèses considérées comme acquises et qui définissent la sécurité alimentaire en termes capitalistes.

Nous ferions mieux de dire la vérité aux sans-pouvoirs

Dire la vérité à ceux qui n'ont aucun pouvoir pourrait être beaucoup plus utile que le cliché qui consiste à « dire la vérité au pouvoir ». Peut-être que les sans-pouvoirs ne savent pas la vérité ou ont une vision confuse de la situation, ce qui contribue à les rendre inactifs, incapable de poursuivre leurs propres intérêts, incapables de voir que d'autres qu'eux partagent les mêmes intérêts.

Nous avons une masse énorme d'approches évaluatives qui promettent la participation, l'émancipation, la transformation… Je fais partie de la communauté évaluative qui fait ces promesses. Je ne dis pas simplement que ce type d'évaluation est meilleur. Mais plutôt, que c'est à nous de nous demander comment l'évaluation pourrait contribuer au bien commun si les pauvres, les sans-abris, ou encore les malades étaient nos clients. Les sans-pouvoirs ont besoin de données probantes et d'analyses qui leur permettront de comprendre [des processus à l'œuvre comme le fait de] blâmer la victime, la privatisation, l'antisyndicalisme et les principes libre-échangistes qui soutiennent les politiques et les idéologies préjudiciables à leur bien-être. Nous avons besoin de plus de vérité pour les impuissants, et non pas pour les puissant-e-s, et nous devons permettre aux impuissant-e-s de parler pour eux-mêmes… Les évaluateurs et les évaluatrices devraient réfléchir à la façon dont nous pourrions faire cela.

Ce sont des idées pour lesquelles je n'ai aucun plan précis, aucune procédure ou processus nécessaires à leur mise en œuvre. Mais, après une vie de travail, je ne suis pas prête à abandonner le potentiel de l'évaluation. Et donc même si notre travail a contribué trop peu au bien commun, a parfois maintenu un *statu quo* préjudiciable, a parfois même contribué à des pratiques sociales néfastes… Je crois que nous pouvons faire mieux. Je crois que nous devrions essayer de faire mieux.

Le regard de Marthe Hurteau

MARTHE HURTEAU

J'ai acquis une formation en évaluation de programme dans le cadre de mon doctorat, ce qui m'a permis d'envisager une carrière axée sur l'enseignement et la recherche. À cela, se sont superposées des activités de consultante qui ont été l'occasion de me familiariser avec des organismes communautaires, comme avec des ministères. Cet éventail d'expériences m'a permis de réaliser qu'il n'y a pas une telle chose que « *one size fits all* ». Ces expériences ont enrichi tant mon enseignement que ma recherche dans le domaine. Je suis professeure à l'Université du Québec à Montréal depuis 2004 et j'ai mis sur pied un programme en évaluation de programme. Je développe ce cheminement et les constats qui s'en dégagent dans mon livre *L'évaluation de programme axée sur le jugement crédible* (2012) et en particulier dans son introduction (p. 1-12).

En ce qui concerne le choix des textes ici présentés, mis à part celui de Bamberger *et al.* (2004), qui date d'ailleurs, il m'apparaît intéressant dans le sens qu'il permet d'aller au-delà du simple processus évaluatif pour aborder des thèmes tels que l'éthique, les modèles, l'apport des contextes et la contribution de la démarche évaluative au bien commun. Cette distanciation du traitement strictement méthodologique me semble aussi essentielle que salutaire au développement de l'évaluation. En effet, si plusieurs écrits ont été publiés sur le sujet au cours des 25 dernières années et ont ainsi contribué à distinguer l'évaluation de la recherche en lui fournissant un cadre de référence qui lui est propre, il est urgent de passer à autre chose. Cependant, pour revenir au choix des textes, s'ils traitent des aspects importants de la pratique, je ne suis pas certaine qu'ils reflètent ce qu'elle est devenue et vers où elle s'en va. C'est le défi des prochaines années d'ailleurs et les thèmes des colloques tant canadiens qu'américains semblent aller dans ce sens : tout est à construire. Il n'en demeure pas moins que j'ai été particulièrement interpelée par le

commentaire de Chelimsky, rapporté par Morris (2010), voulant qu'il soit nécessaire de faire preuve de courage face aux défis auxquels l'évaluateur ou l'évaluatrice est confronté-e. Si cette citation était pertinente en 1995, elle m'apparaît encore plus d'actualité.

Pourquoi faut-il que les évaluateurs et les évaluatrices soient encore plus courageux et courageuses? Il est évident que nous devons considérer le contexte dans lequel nous vivons et je pense que cette citation en fait foi :

> Nicolletta Stame (2018) introduit son propos de la façon suivante : *Il est nécessaire de renforcer la contribution éthique des évaluateurs dans le monde tumultueux actuel*. Nous ne pouvons faire autrement que de souscrire à ce constat alors que la situation actuelle continue à se dégrader. En effet, de récents événements tels que la gestion mondiale de la pandémie de COVID-19 et la distribution des vaccins, les élections américaines et les changements climatiques nous ont amenés à réfléchir sur le monde dans lequel nous vivons. Dans son livre *21 Lessons for the 21st Century*, Yuval Noah Harari (2018) note que les humains appuient largement leurs jugements et leurs décisions importantes sur leurs émotions et leurs préjugés. Le manque d'information et l'incapacité à distinguer le faux du vrai (pour ne citer que les fausses nouvelles), incitent le subconscient à générer une analyse réductionniste des problèmes et à envisager des solutions aussi faciles que rassurantes. De plus en plus, le monde se divise en « bon » et « mauvais » et la réalité se résume à quelques faits ou histoires touchantes qui en viennent à incarner La Vérité. Celle-ci se décline en différentes versions, ce qui est d'autant plus troublant que la responsabilité de l'évaluateur consiste à générer un jugement sain, éthique, s'appuyant sur des faits, et ce, dans le but d'éclairer la prise de décision. De plus, l'évaluateur est parfois appelé à adopter une posture éthique et à la soutenir. Cela se révèle encore plus vrai en temps de crise et de catastrophe (Jakubik, 2020) (traduction libre, Hurteau et Gagnon, soumis)

Mais il y a plus. Des voix s'élèvent au niveau de communautés, pour ne citer que les Afro-Américain-e-s et les Premières Nations, pour avoir un traitement qu'elles jugent équitable et cela se traduit au niveau de leur participation à des activités professionnelles ainsi que de leur acceptation à collaborer avec des personnes à l'extérieur de leur communauté.

Cela m'amène à aborder la thématique des parties prenantes avec lesquelles les évaluateurs et les évaluatrices sont appelé-e-s à transiger. Comme le soutient M-P Marchand (2020) dans le cadre de sa thèse de doctorat, elles sont de plus en plus éduquées. En effet, elles ont acquis des connaissances et de l'expérience à travers les années. Elles sont ainsi en mesure de confronter les évaluateurs et les évaluatrices tant au niveau des choix stratégiques qu'au niveau des choix méthodologiques proposés. Il est surtout important de souligner l'enjeu qui se dégage de cette recherche. En effet, pour les parties prenantes, l'évaluation sera crédible si elle est utile, sous-entendant utile à obtenir du financement ou à ne pas le perdre. En effet, comme de plus en plus d'évaluations s'inscrivent dans une démarche de reddition de comptes, la survie des programmes est directement reliée à leur capacité à générer les résultats annoncés (efficacité du programme). Les évaluateurs et les évaluatrices sont interpelé-é-s pour livrer une évaluation qui va dans ce sens et ils et elles sont parfois confronté-e-s à de fortes pressions lorsque les résultats ne sont pas au rendez-vous. Ils et elles interpellent à leur tour les associations professionnelles et les théoriciens et les théoriciennes pour les soutenir dans cette situation, mais le code d'éthique ou tout document du genre n'est pas encore écrit. Et s'il existait un jour, il sera aussi volumineux que la Bible, et encore, afin de prendre en considération toutes les situations qui peuvent se présenter.

En effet, j'ai été aux premières loges pour recevoir ces demandes dans le cadre de ma participation aux travaux de révision des lignes directrices en éthique au sein de la Société canadienne d'évaluation (SCÉ). Alors que les évaluateurs et les évaluatrices, et surtout les novices, requièrent des balises claires, une éthique axée sur la déontologie (fais ce que dois), les travaux réalisés par le groupe de travail nous amènent plutôt à

recommander – compte tenu de ce qui vient d'être mentionné – une approche éthique axée sur la sagesse pratique : l'écart est grand! Pourquoi? Parce qu'elle va dans le sens opposé aux demandes, la sagesse pratique visant à faire la bonne chose dans un contexte particulier afin d'atteindre un but ou, dit différemment, à régler un problème (House, 2015). La SCÉ semble s'orienter dans ce sens et nous sommes conscients que cela devra s'accompagner d'activités de formation et de soutien aux membres.

Pourquoi ce choix qui semble irrationnel puisqu'à l'encontre des demandes ? Des théoriciens tels que Ernest House (2015, 2018), Thomas Schwandt (2015, 2017a, 2017b, 2018), Arnold Love (2018) et James McDavid (2019) sont d'avis qu'il n'existe pas un « meilleur » modèle d'éthique et que le jugement professionnel implique d'être conscient-e des divers types de pressions éthiques qui peuvent être en jeu dans un contexte précis et d'être capable de naviguer de manière réfléchie dans la situation. Dans cette perspective, la sagesse pratique constitue une option valable qui rehausse la légitimité et la valeur de la démarche au titre du bien social, tout en favorisant l'autonomie professionnelle.

Plusieurs types d'intervenants, pour ne citer que les médecins, les infirmiers et infirmières, les gestionnaires, les agent-e-s de première ligne, etc. – ont compris qu'une pratique professionnelle efficace et empreinte d'éthique va bien au-delà de la simple application de méthodes et de techniques. En effet, ils et elles doivent improviser, prendre en considération des objectifs contradictoires et interpréter des règles à la lumière des particularités de chaque contexte. Mais avant toute chose, ils et elles doivent souvent faire preuve de perspicacité et de courage, comme l'indique Chelimski.

Un nombre impressionnant de publications scientifiques documentent la contribution de la sagesse pratique à l'acte professionnel dans de nombreux métiers (Hurteau et Gagnon, soumis). J'illustrerai mon propos par un exemple qui n'est pas directement relié à l'évaluation de programme – parce que cela peut devenir délicat, mais d'actualité. En

décembre dernier, le Canada ne disposait que de très peu de vaccins COVID-19. Les autorités du Québec ont décidé, compte tenu de la situation, d'administrer une première dose à plus de personnes plutôt que de respecter le protocole suggéré et d'espacer les doses de 21 jours. Elles se basaient sur le principe de la science des vaccins qui suggère de permettre au corps de développer ses anticorps, ainsi que sur des valeurs profondes : sauver le plus de vies possibles. Cette décision s'est appuyée sur la science (des vaccins), sur une intuition (et si c'était comme les autres vaccins), sur une grande dose de courage parce que cette décision a été fortement contestée et… sur un peu de chance. Il s'est avéré que ce choix fut judicieux : le Québec a traversé la troisième vague sans trop de problèmes et bien des provinces canadiennes, comme des pays, ont finalement adopté cette posture. Nous comprenons que la sagesse pratique est mise à contribution parce qu'il y avait un problème. En effet, contrairement à la situation aux États-Unis, l'approvisionnement était limité au Canada. Cela s'avérait la meilleure décision dans les circonstances. Les évaluateurs et les évaluatrices sont constamment confronté-e-s à ce genre de décision : la meilleure dans les circonstances. En effet, la pratique évaluative éclate, tout en devenant de plus en plus politisée : le « *one size fits all* » n'est plus de mise! La situation est problématique : manque de ressources et de temps; des client-e-s qui interpellent les choix stratégiques et méthodologiques ainsi que les résultats générés. Bref, de nombreux enjeux dont la solution ne peut être clairement dictée dans des normes professionnelles ou des lignes directrices en éthique.J'ai aussi participé à la révision du référentiel de compétences élaboré par le SCÉ dans le cadre de son programme d'accréditation des évaluateurs et des évaluatrices, et le changement le plus essentiel réside dans la place qui est maintenant accordée à la pensée réflexive. En effet, elle s'ajoute aux compétences requises pour effectuer une évaluation – connaissance des normes et compétences techniques, interpersonnelles, de gestion et situationnelles. Pourquoi? Parce que l'évaluateur ou l'évaluatrice doit réfléchir à ce qu'il ou elle fait, et non pas seulement mettre en application une méthodologie, et qu'il ou elle doit revenir sur ses expériences afin d'en tirer les leçons qui s'imposent.Ainsi,

quelques conseils pour les évaluateurs et les évaluatrices, qui sont d'ailleurs abordés par Love (2018) : prendre connaissance des récents écrits; mettre sur pied des groupes de réflexion qui leur permettent de revenir sur leurs « bons et moins bons » coups et d'envisager les apprentissages qu'ils et elles peuvent faire. Finalement : demeurer humble parce que la tâche est complexe.

Bibliographie

Bamberger, Michael, Jim Rugh, Mary Church et Lucia Fort. 2004. « Shoestring evaluation: Designing impact evaluations under budget, time, and data constraints ». *American Journal of Evaluation* 25(1) : 5-37. doi : https://doi.org/10.1177%2F109821400402500102.

House, Ernest R. 2015. *Evaluating: Values, Biases, and Practical Wisdom*. Charlotte: Information Age Publishing Inc.

House, Ernest R. 2018. « L'apport de la sagesse pratique ». in *L'évaluation de programme axée sur la rencontre des acteurs*, édité par M. Hurteau, I. Bourgeois et S. Houle. Québec : Presses de l'Université du Québec, p. 25-34.

Hurteau, Marthe, et Caroline Gagnon. Soumis. « Can Practical Wisdom Contribute to an Ethical Evaluation Practice? ». *Evaluation, The International Journal of Theory, Research and Practice.*

Hurteau, Marthe, Sylvain Houle et François Guillemette, éd. 2012. *L'évaluation de programme axée sur le jugement crédible*. Québec : Presses de l'Université du Québec.

Love, Arnold. 2018. « De la sagesse pratique à une pratique empreinte de sagesse ». in *L'évaluation de programme axée sur la rencontre des acteurs*. Québec : Presses de l'Université du Québec, p. 71-91.

Marchand, Marie-Pier. 2020. « Vers une meilleure contextualisation de l'évaluation crédible : identification des facteurs d'influence de la crédibilité à partir d'expériences évaluatives de parties prenantes ». Thèse ou essai doctoral accepté, Université du Québec, Montréal.

McDavid, James C., Irene Huse et Laura R. L. Hawthorn. 2019. *Program Evaluation and Performance measurement. An Introduction to Practice*. 3e éd. New York: Sage Publications.

Moris, Michael. 2011. « The good, the bad and the evaluation: 25 years of AJE Ethics ». *American Journal of Evaluation* 32(1) : 134-151

Schwandt, Thomas A. 2015. *Evaluation Foundations Revisited: Cultivating a Life of the Mind for Practice*. Stanford: Stanford University Press.

Schwandt, Thomas A. 2017a. « Becoming better evaluators through reflective practice: adventures & insights from diverse evaluators working in academic, government, & consulting contexts ». Washington.

Schwandt, Thomas A. 2017b. « Professionalization, Ethics, and Fidelity to an Evaluation Ethos ». *American Journal of Evaluation* 20(10) : 1-8. doi : https://doi.org/10.1177/1098214017728578.

Schwandt, Thomas A. 2018. « Evaluative thinking as a collaborative social practice: The case of boundary judgment making ». *New Directions for Evaluation* 2018(158) : 125-37. doi : https://doi.org/10.1002/ev.20318.

III. COMMENT JUGER DE LA VALEUR DES INTERVENTIONS?

Introduction : évaluer en fonction de quelles valeurs?

THOMAS DELAHAIS, AGATHE DEVAUX-SPATARAKIS, ANNE REVILLARD ET VALÉRY RIDDE

Étymologiquement, évaluer c'est déterminer la valeur. C'est là que les ennuis commencent. En effet, le terme « valeur » est ambigu : la valeur peut désigner, d'une part, le prix d'une chose, « ce qui coûte » — et aussi, d'autre part, l'importance qu'elle revêt — « ce qui compte ». C'est cette seconde définition qui nous intéresse dans cet ouvrage, qu'il s'agisse d'une qualité recherchée, de finalités communes, de principes politiques, éthiques ou moraux, etc., quoiqu'elle ne suffise pas à lever complètement l'ambiguïté. Même en considérant qu'une intervention a de la valeur, on peut vouloir en avoir pour son argent — c'est la logique *value for money* que l'on retrouve dans les approches d'évaluation économique telles que l'analyse coût-avantage ou coût-efficacité (King, 2017).

Il faut tout de suite écarter l'idée que l'évaluation serait un jugement de valeur, au sens de l'appréciation purement subjective et spontanée d'une situation. Dans tous les cas, l'évaluation s'appuie sur un processus structuré visant à établir des faits relatifs à l'intervention évaluée et à ses conséquences, ces faits permettant la construction progressive d'un jugement. On parle ici d'un jugement sur la valeur, c'est-à-dire selon lequel ces faits sont analysés et interprétés à l'aune de « ce qui compte » dans le contexte de l'évaluation. Évidemment, « ce qui coûte » est aussi une dimension utile, mais dont il ne sera pas question dans cette partie. Le processus d'élaboration du jugement a déjà été présenté en français par Bernard Perret (2012). La discussion porte ici sur la nécessité d'un tel jugement, ainsi que sur la légitimité à le porter, et sur la façon d'établir les valeurs en fonction desquelles juger, en particulier au regard des conflits de valeur et des rapports de domination qui traversent nos sociétés. Ce sont ces différents débats que nous allons explorer.

Fondements : l'évaluation, activité descriptive ou jugement sur la valeur?

L'évaluation comme activité descriptive

Pour tout un courant de pensée, l'évaluation est d'abord considérée comme une activité technique (de structuration, de collecte, d'analyse, etc.) qui vise à poser des hypothèses relatives à une intervention, à les examiner en mobilisant des outils des sciences sociales, et à vérifier ainsi que les résultats sont conformes à ce qui est attendu (e.g. Rossi, Lipsey, et Freeman, 2003). Dans ce processus, les valeurs n'ont pas leur place. Les hypothèses sont basées sur les objectifs affichés d'une intervention, ou bien tirées de théories issues des sciences sociales. L'enjeu est alors d'élaborer la meilleure méthode, la plus rigoureuse, pour les tester.

Ce n'est pas que les soutiens de ce point de vue ne s'intéressent pas à la notion de « bien » ou de « bon », mais plutôt qu'ils et elles considèrent devoir tenir à distance de leurs travaux toute idée de valeur, pour aboutir à une description objective de « ce qui fonctionne » (ou dysfonctionne) : ce travail de description en tant que tel passe par une démarche de mise à distance des valeurs. Seulement, alors, la société pourra choisir les bonnes solutions au service d'une vie meilleure, comme l'exprime Donald Campbell dans son projet de société expérimentale (1969). *Mutatis mutandis*, telle est la vision que l'on retrouve aujourd'hui dans le courant de la politique basée sur des données probantes (*evidence-based policy*). C'est en multipliant les évaluations et en agrégeant leurs résultats selon ces principes qu'émergeront les interventions capables de résoudre les problèmes de nos sociétés, plaide Howard White (2019) **(texte 1)**.

… ou une activité de jugement ?

À côté de ce courant s'en dessine un autre avec Michael Scriven **(texte 2)**. Pour lui, justement, ce qui distingue l'évaluation, c'est l'attention portée au jugement. Scriven rappelle que l'évaluation est une activité fondamentale de l'être humain, qui lui permet de se faire un avis et faire des choix. Ce processus, cependant, est largement intériorisé. L'évaluation appliquée aux programmes s'attache, elle, à définir explicitement ce qui est « bien » pour élaborer un jugement, grâce à un ensemble de critères et à un « niveau de performance » (*performance standards*) à atteindre. Alors seulement intervient la mesure, et pour finir, la production d'un jugement[1].

Le processus qui va des faits à la détermination de la valeur est ce que Michael Scriven appelle *logique de l'évaluation*. C'est la définition reprise par l'Association américaine d'évaluation : « L'évaluation est le processus par lequel on détermine le mérite, l'intérêt (*worth*) et l'importance (*significance*) des choses » (1991). Là, le *mérite* renvoie aux qualités intrinsèques d'une intervention, l'*intérêt* à son apport dans un contexte précis, et l'*importance* à un jugement plus global sur l'intervention combinant mérite et intérêt — le tout constituant sa valeur (*value*). Pour Scriven, il est possible de déterminer assez facilement quels sont les éléments qui fondent le mérite d'une intervention. Dans une de ses analogies favorites avec l'évaluation de produits par les associations de consommateurs, il dit : « Si vous savez ce qu'est une montre, vous savez que le mérite d'une montre renvoie à son exactitude, sa lisibilité et sa durabilité; et si vous savez cela, vous savez comment (de prime abord) établir des conclusions évaluatives à partir d'éléments factuels sur le mérite comparatif de différentes montres » (Scriven, 1991 : 217).

1. Nous renvoyons sur ce processus de formation du jugement au texte de Bernard Perret, en français, « Construire un jugement » (2012).

Dans cette logique, on peut alors vouloir établir un ensemble de critères s'appliquant à toutes les interventions. En 1991, le Comité d'Aide au Développement (CAD) de l'Organisation de Coopération et de Développement Économiques (OCDE) en a dégagé cinq (pertinence, efficacité, efficience, impact et durabilité), qui se sont largement imposés dans le monde du développement[2]. Mais ces cinq-là n'épuisent pas la liste de ceux qui peuvent être appliqués aux interventions évaluées. Daniel L. Stufflebeam, par exemple, propose une liste élargie de valeurs et de critères à prendre en compte dans une évaluation, incluant notamment des valeurs sociétales (l'équité, la liberté, la citoyenneté, mais aussi le respect des lois ou la défense nationale, **texte 3**). Hassall et coll., de leur côté, identifient cinq grandes catégories de valeurs qui peuvent affecter la détermination de « ce qui compte » ou « ce qui est bien » dans une évaluation : personnelles, sociales, politiques, professionnelles et épistémiques (2020).

Bien entendu, les deux processus, celui qui consiste à décrire et celui qui consiste à juger, ne sont pas antinomiques, ils sont même complémentaires et peuvent être concomitants. Robert Stake explique ainsi que « pour être totalement compris, [une intervention] doit être totalement décrite et totalement jugée » (1977, cité par Gullickson, 2020 : 2). Précisons ici que les deux processus se déroulent simultanément : autrement dit, il ne s'agit pas simplement de bien décrire pour ensuite bien juger. Pour rester dans la logique de l'évaluation de Scriven, il faut d'abord avoir délimité le périmètre d'évaluation, et avoir compris l'intervention afin de s'accorder sur des critères de jugement (« qu'est-ce qui serait bien? ») et des niveaux de performance. La collecte d'information est organisée de façon à vérifier les critères, elle est donc informée par les valeurs; et les jugements portés en fin d'évaluation s'appuient par les éléments descriptifs recueillis.

2. En 2019, un critère de cohérence a été ajouté et les définitions ont été revues. Voir sur le site de l'OCDE : https://www.oecd.org/fr/cad/evaluation/criteres-cad-evaluation.htm

Controverses : Qui doit définir les critères de jugement?

L'évaluateur/-trice à partir des visions des parties prenantes?

Pour Scriven, si on connaît une chose, on sait comment la juger. Il réfute la nécessité de connaître les valeurs que les uns et les autres attribuent à un objet, y compris à une intervention publique, pour pouvoir l'évaluer (voir **texte 2**). Cette position devient cependant sujette à contestation dans les années 1970. Comme le dit Robert Stake :

> Une œuvre d'art n'a pas une valeur unique. Une intervention n'a pas une valeur unique. Pourtant ils ont tous les deux de la valeur. La valeur d'un programme d'éducation artistique sera différente pour des personnes différentes, selon les usages prévus… S'il y a un consensus sur les valeurs… il doit être découvert. (1975, cité par Abma et Stake, 2001 : 9)

Ce changement de conception ouvre des perspectives nouvelles sur l'évaluation et son rôle. Car alors, s'il n'y a pas « une bonne façon de juger », il faut élaborer un processus visant à faire émerger les valeurs et à les rendre opérationnelles dans le contexte de l'évaluation. Pour Stake, il s'agit d'un processus hybride d'écoute d'une pluralité de parties prenantes, et aussi de respect de la diversité des points de vue. À la fin de ce processus, cependant, comme chez Scriven, c'est encore à l'évaluateur ou à l'évaluatrice de transformer les valeurs en critères de jugement et en niveaux de performance : « Ce qui est mauvais est mauvais et ce qui est bon est bon et c'est le travail des évaluateurs de décider lequel est lequel » (Scriven, 1986, cité par Alkin et Christie, 2004 : 32).

… ou les parties prenantes elles-mêmes?

Or, dans un monde dans lequel l'évaluation est essentiellement une activité de marché (Lemire et coll., 2018 ; et pour la France, Matyjasik, 2010), on peut se poser la question de savoir si ce ne sont pas les commanditaires qui imposent leurs valeurs, à travers leurs objectifs, ou les questions évaluatives posées. C'est la « tendance au managérialisme » que reprochent Egon Guba et Yvonne Lincoln aux évaluateurs et aux évaluatrices des générations précédentes (1989), c'est-à-dire la propension à accorder plus d'importance aux attentes des gestionnaires ou aux financeurs des interventions, au détriment des autres parties.

Certes, Scriven appelle à se détacher des objectifs et à juger de leur pertinence avant d'en faire des critères d'évaluation, mais il n'est pas sûr que les évaluateurs et les évaluatrices aient toujours la hauteur de vue qu'il leur prête, ou d'ailleurs une marge de manœuvre suffisante pour effectuer ce travail de mise à distance. Celles et ceux qui évaluent ont bien sûr des valeurs qui leur sont propres. Mais, sans en être toujours conscient-e-s, n'est-il pas possible que ces valeurs soient les mêmes que celles de leurs commanditaires? Et cela n'affecte-t-il pas leurs jugements? Finalement, même une évaluation dite indépendante, se targuant de « dire la vérité au pouvoir », n'est pas à l'abri de finir par dire « la vérité *du* pouvoir » (Mathison, 2017).

Une autre approche consisterait alors à aider l'ensemble des parties prenantes à exprimer leurs valeurs, et les aider à formuler un jugement commun. C'est ce que font Guba et Lincoln **(texte 4)**, avec leur évaluation « de 4e génération ». Mais il ne s'agit pas que de participation. Leur approche s'éloigne également de la logique de l'évaluation de Scriven avec ses critères et ses niveaux de performance, pour se rapprocher d'une démarche d'investigation systématique des affirmations des parties prenantes (affirmations empreintes de valeurs) dans une recherche progressive de consensus.

Pour Guba et Lincoln, chacun-e des protagonistes d'une intervention évaluée a sa propre expérience de la réalité, et aucun-e ne peut prétendre à la vérité. Ces « constructions » se reflètent dans les affirmations, les jugements, les enjeux, les points d'attention qu'ils portent *a priori* sur l'intervention évaluée. Le but de la démarche évaluative est ainsi d'identifier et de réunir toutes les parties prenantes de l'intervention, puis de collecter ces constructions et de s'accorder sur celles qui seront testées lors de l'évaluation. Dans ce processus, les désaccords sont au cœur de la démarche. Ils sont portés au centre de l'attention, font l'objet d'une collecte d'information et d'un processus de négociation *ad hoc* : dès qu'une affirmation est consensuelle, elle est acquise, et l'évaluation se concentre sur les points conflictuels suivants. En ce sens, une évaluation constructiviste ne se termine jamais, elle est mise en sommeil jusqu'au moment où une nouvelle occasion favorable se produit.

L'approche qui est décrite ici est particulièrement exigeante et il est peu d'exemples de sa mise en œuvre (Lay et Papadopoulos, 2007). Mais elle trace la voie pour des évaluations prenant sérieusement en compte les parties prenantes et leurs valeurs, comme l'ont montré Donna Mertens et Jennifer Greene. Cette dernière s'inscrit dans l'héritage de l'évaluation démocratique délibérative de Ernst House et Kenneth Howe (1999), mais en accordant une place essentielle à la détermination des valeurs des parties prenantes et à leur participation dans ce processus. Elle en donne à voir la dimension très pratique, dans un texte sur l'évaluation « engagée dans les valeurs » (2005) **(texte 5)**.

Perspectives : en défense de valeurs spécifiques?

À l'affirmation, fermement établie dans les années 1990, que l'évaluation peut difficilement être neutre en ce qui concerne les valeurs, on peut donc répondre en explicitant les valeurs des un-e-s et des autres. On peut toutefois déplorer qu'il soit souvent difficile d'inclure toute la variété

des points de vue dans une évaluation. À reconnaître une multiplicité de valeurs, on risque aussi de mettre toutes les perspectives au même niveau. Faut-il dès lors privilégier certaines valeurs en particulier, ou bien évaluer en fonction de valeurs transcendant celles qui sont exprimées dans les politiques publiques ou dans les points de vue de chaque partie prenante prise séparément?

L'évaluation au service des valeurs des dominé-e-s

Ernest House a très tôt montré la voie dans la première approche en s'inspirant de la théorie de la justice sociale de John Rawls. Pour Rawls, les inégalités ne peuvent être acceptées dans une société que si elles permettent d'améliorer la situation des plus désavantagé-e-s. C'est ce qu'applique House à l'évaluation. Pour lui, les évaluateurs et les évaluatrices ont une responsabilité morale à considérer les conséquences de leur activité pour la société dans son ensemble, et en particulier pour les publics marginalisés (Christie et Alkin, 2008). Il les enjoint ainsi à réfléchir à leurs propres valeurs et à celles qu'ils et elles veulent soutenir dans leur activité. Il est pour lui acceptable de donner plus d'importance aux valeurs des personnes qui n'ont jamais voix au chapitre si cela permet d'améliorer leur situation. De même, les évaluateurs et les évaluatrices ont un rôle à jouer pour s'opposer au culte du rapport coût-efficacité ou à d'autres théories affectant les plus pauvres (House, 2004). Au bout du compte, il ne s'agit plus de penser en termes de bon ou de mauvais, comme le faisait Scriven, mais en termes de *right, fair or just* : ce qui est approprié, équitable ou correct.

C'est cette même logique que suivent globalement les évaluateurs et les évaluatrices transformationnel-le-s, dans la foulée de Jennifer Greene et de Donna Mertens. L'évaluation féministe est un exemple d'approche transformationnelle. Elle exprime clairement ses valeurs, assume l'idée que les rapports de genre mènent à des situations d'injustice sociale, et

propose d'utiliser le prisme du genre pour repérer et expliquer comment les injustices se produisent et sont maintenues ou renforcées par les interventions évaluées. Sur ce sujet nous renvoyons à l'article en français de Donna Podems : *Rendre l'évaluation féministe praticable* (2018).

L'évaluation féministe amène l'équipe d'évaluation et les parties prenantes impliquées dans l'évaluation à s'interroger sur leurs propres valeurs implicites. Il en est de même pour l'évaluation attentive aux différences culturelles (*culturally responsive evaluation*, CRE). La CRE part du principe que les évaluations sont inscrites dans une culture et que « des valeurs et des croyances liées à la culture sont au cœur de tout effort évaluatif » (Hood, Hopson, et Kirkhart, 2015 : 283), avec une attention toute particulière portée aux groupes sociaux ou racisés qui ont été ou sont encore marginalisés. L'évaluation attentive à la culture interroge la façon dont une intervention respecte la culture des groupes concernés dans ses intentions, ses hypothèses sous-jacentes et son fonctionnement. L'évaluation indigène **(texte 6)** va au bout de cette logique en amenant des systèmes de valeurs complètement différents de ceux des pays du Nord. Par exemple, l'Ubuntu appelle à prendre en compte les interactions dans la communauté comme une composante de l'identité, ou encore les conséquences de nos actes sur le vivant et le non vivant (Chilisa *et al.*, 2016). Que serait une évaluation prenant pleinement en compte ces valeurs? Avec, à terme, la question : faut-il que seules des personnes autochtones puissent mener des évaluations en contexte autochtone? (Wehipeihana, 2019).

Prendre en compte des valeurs supérieures?

Pour finir, un autre point de vue consisterait à embrasser un ensemble de valeurs jugées supérieures. House traçait déjà cette voie en appelant à évaluer les interventions en termes de justice sociale. Dans son sillage, un mouvement s'est constitué pour généraliser le recours à l'équité comme

critère d'évaluation. En France, les approches d'utilité sociale (Offredi et Ravoux, 2010) placent au premier plan des valeurs telles que la solidarité, le bien-être individuel et social, le lien social, les biens publics, la cohésion sociale et la reconnaissance. Plus récemment des approches visant à réaffirmer le rôle de l'évaluation en soutien à l'intérêt général (Picciotto 2015), ou à assurer un futur durable à l'humanité à l'ère de l'anthropocène se sont fait jour (Blue Marble Evaluation, Patton, 2020). Le recours à ces valeurs supérieures donne du sens au processus évaluatif — mais il est aussi dans certains cas un moyen de dialoguer sur ce qui compte. En effet, pour faire du bien-être, par exemple, un critère d'évaluation, encore faut-il s'accorder sur ce qu'il recouvre, ce qui rend nécessaire un processus de dialogue et de délibération (Offredi et Laffut, 2013). Le recours à des valeurs supérieures comme critères d'évaluation est ainsi une façon de réaffirmer la dimension politique de l'évaluation.

Bibliographie

Abma, Tineke A., et Robert E. Stake. 2001. « Stake's Responsive Evaluation: Core Ideas and Evolution ». *New Directions for Evaluation* 2001(92) : 7. doi : 10.1002/ev.31.

Alkin, M. C., et Christina A. Christie. 2004. « An evaluation theory tree ». in *Evaluation Roots*. Thousand Oaks: Sage.

Campbell, Donald T. 1969. « Reforms as experiments ». *American Psychologist* 24(4) : 409-29. doi: https://doi.org/10.1037/h0027982.

Chilisa, Bagele, Thenjiwe Emily Major, Michael Gaotlhobogwe et Hildah Mokgolodi. 2016. « Decolonizing and Indigenizing Evaluation Practice in Africa : Toward African Relational Evaluation Approaches ». *Canadian Journal of Program Evaluation* 30(3) : 313-28. doi: 10.3138/cjpe.30.3.05.

Christie, Christina A., et Marvin C. Alkin. 2008. « Evaluation Theory Tree Re-Examined ». *Studies in Educational Evaluation* 34(3) : 131-35. doi : 10.1016/j.stueduc.2008.07.001.

Greene, Jennifer C. 2005. « A Value-Engaged Approach for Evaluating the Bunche-Da Vinci Learning Academy ». *New Directions for Evaluation* 2005(106) : 27-45. doi : 10.1002/ev.150.

Guba, Egon G., et Yvonna S. Lincoln. 1989. *Fourth Generation Evaluation*. SAGE Publications Ltd.

Gullickson, Amy M. 2020. « The Whole Elephant: Defining Evaluation ». *Evaluation and Program Planning* 79 : 101787. doi : 10.1016/j.evalprogplan.2020.101787.

Hassall, Keryn, Amy M. Gullickson, Ayesha S. Boyce et Kelly Hannum. 2020. « Editorial ». *Evaluation Journal of Australasia* 20(2) : 63-67. doi : https://doi.org/10.1177/1035719X20931250.

Hood, Stafford, Rodney K. Hopson et Karen E. Kirkhart. 2015. « Culturally responsive evaluation ». *Handbook of practical program evaluation* 281.

House, Ernest, et Kenneth R. Howe. 1999. *Values in Evaluation and Social Research*. SAGE Publications.

House, Ernest R. 2004. « The Role of the Evaluator in a Political World ». *Canadian Journal of Program Evaluation* 19(2) : 16.

King, Julian. 2017. « Using Economic Methods Evaluatively ». *American Journal of Evaluation* 38(1) : 101-13. doi : 10.1177/1098214016641211.

Lay, Margaret, et Irena Papadopoulos. 2007. « An Exploration of Fourth Generation Evaluation in Practice ». *Evaluation* 13(4) : 495-504. doi : 10.1177/1356389007082135.

Lemire, Sebastian, Steffen Bohni Nielsen et Christina A. Christie. 2018. « Toward Understanding the Evaluation Market and Its Industry-Advancing a Research Agenda ». *New Directions for Evaluation* 2018(160) : 145-63. doi : 10.1002/ev.20339.

Mathison, Sandra. 2018 [2017]. « Does Evaluation Contribute to the Public Good? ». *Evaluation* 24(1) : 113-19.

Matyjasik, Nicolas. 2010. « L'évaluation des politiques publiques dans une France décentralisée. Institutions, marché et professionnels ». Université de Bordeaux; Université Montesquieu-Bordeaux IV; Institut d'études politiques de Bordeaux; SPIRIT.

Offredi, Claudine, et Michel Laffut. 2013. « Le bien-être peut-il être un critère d'évaluation de l'action publique ? » *Revue française d'administration publique* (148) : 1003-16.

Offredi, Claudine, et Françoise Ravoux. 2010. *La notion d'utilité sociale au défi de son identité dans l'évaluation des politiques publiques*. Paris : L'Harmattan.

Patton, Michael Quinn. 2020. *Blue marble evaluation: premises and principles*. New York: The Guilford Press.

Perret, Bernard. 2012. « La construction d'un jugement ». in V. Ridde et C. Dagenais. *Approches et pratiques en évaluation de programmes*. Les Presses de l'Université de Montréal, p. 53-70.

Picciotto, Robert. 2015. « Democratic Evaluation for the 21st Century ». *Evaluation* 21(2) : 150-66. doi : 10.1177/1356389015577511.

Podems, Donna. 2018. « Rendre l'évaluation féministe praticable ». *eVALUation Matters*.

Rossi, Peter H., Mark W. Lipsey et Howard E. Freeman. 2003. *Evaluation: a systematic approach*. 7th ed. Thousand Oaks, CA: Sage.

Scriven, Michael. 1991. *Evaluation thesaurus*. 4th ed. Newbury Park, Calif: Sage Publications.

Wehipeihana, Nan. 2019. « Increasing Cultural Competence in Support of Indigenous-Led Evaluation: A Necessary Step toward Indigenous-Led Evaluation ». *Canadian Journal of Program Evaluation* 34(2). doi : 10.3138/cjpe.68444.

White, Howard. 2019. « The Twenty-First Century Experimenting Society: The Four Waves of the Evidence Revolution ». *Palgrave Communications* 5(1) : 47. doi : 10.1057/s41599-019-0253-6.

1. La société expérimentale du XXIe siècle : les quatre vagues de la révolution de la preuve

HOWARD WHITE

[Traduit de : White, Howard. 2019. « The twenty-first century experimenting society : The four waves of the evidence revolution ». *Palgrave Communications*, 5(1) : 1-7 (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Palgrave Communications.]

Près des deux tiers des écoles d'Angleterre utilisent les données issues de revues systématiques pour décider de l'affectation des ressources scolaires et planifier les activités en classe. L'ONG américaine de développement, *International Rescue Committee* (IRC), s'est engagée à faire en sorte que tous ses programmes soient fondés sur des données probantes ou produisent des données probantes d'ici 2020. En décembre 2018, le Congrès des États-Unis a adopté la loi sur l'élaboration de politiques fondées sur des données probantes. Au Royaume-Uni et aux États-Unis, le mouvement « *What Works* » [Qu'est-ce qui marche/Ce qui marche] fournit des éléments de preuve quant à l'efficacité des interventions visant à améliorer l'apprentissage, à réduire la maltraitance des enfants et le nombre de sans-abris, à lutter contre la criminalité et à améliorer le bien-être.

Avons-nous atteint la vision de Donald Campbell d'une société expérimentale[1]? C'est-à-dire une société dans laquelle les choix en matière de politique sociale sont fondés sur des données tirées de recherches de haute qualité – « test par ingénierie sociale au cas par cas » (cité par Campbell, 1988). Alors que ce type d'expérimentation existe depuis les années 1930 – principalement aux États-Unis – un changement radical s'est produit ces dernières années, en partie grâce au mouvement *What Works*. Il est juste de parler d'une révolution de la preuve. Cette révolution a commencé il y a soixante-dix ans dans le secteur de la santé avec une médecine fondée sur des données probantes (Oliver et Pearce, 2017). Dans d'autres secteurs, tels que le développement international, l'éducation et la protection sociale, la révolution de la preuve a globalement suivi les quatre vagues décrites dans le présent commentaire. Le récit ci-après décrit fidèlement l'expérience dans le développement international. Il est axé sur ce qui a été fait, et peut être fait, pour soutenir l'utilisation des données probantes dans la prise de décision. Bien entendu, de nombreux autres facteurs influencent la prise de décision. En fin de compte, la politique est un processus politique (voir, par exemple, (Cairney, 2016 et Parkhurstm 2017), mais ces questions ne sont pas examinées plus avant ici. […]

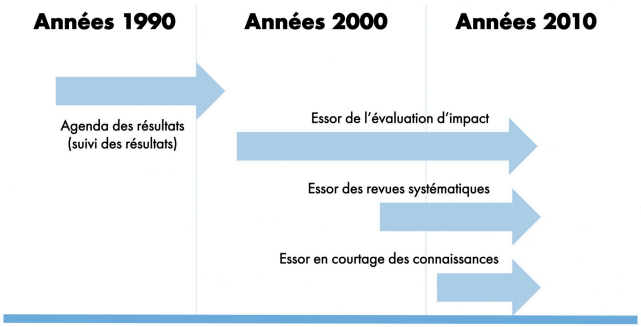
Le présent article porte tout autant sur la demande de données probantes que sur l'offre. Il décrit des initiatives prises par les commanditaires d'activités de recherche et les utilisateurs et les utilisatrices, et pas seulement les producteurs et les productrices, ainsi que des façons de soutenir la demande. Il s'intéresse en particulier à l'institutionnalisation de l'utilisation de données probantes.

Cette institutionnalisation peut être observée dans les quatre vagues de la révolution de la preuve : (1) l'agenda axé sur les résultats, (2) les évaluations d'impact, (3) les revues systématiques et (4) le courtage en

1. La vision de Campbell pour la société expérimentale est exposée dans Campbell (1969). La contribution complète de Campbell à un éventail de disciplines peut être lue dans Boruch (2019).

connaissances (voir figure 1). Cet article décrit l'évolution de la révolution à travers ces quatre vagues avec des exemples venant du monde entier, mais surtout de ma propre expérience dans le développement international.

Figure 1 : Les quatre vagues de la révolution de la preuve



La première vague : l'agenda des résultats, 1990

La révolution de la preuve est apparue dans le cadre de la nouvelle gestion publique qui s'est imposée dans les pays anglophones et scandinaves dans les années 1990. Parmi les faits marquants, citons le *Government Performance and Results Act* (GPRA) [Loi sur la performance et les résultats du gouvernement] de 1993 aux États-Unis et le Livre blanc de 1999 sur la modernisation du gouvernement au Royaume-Uni. La nouvelle gestion publique oblige les organismes gouvernementaux à rendre compte de leurs performances, telles qu'elles ressortent des tendances

(résultats) de haut niveau comme le chômage, la pauvreté, etc. Ce passage à une approche axée sur les résultats a été une réalisation importante, car la performance avait été évaluée jusqu'à présent simplement en fonction des intrants tels que les sommes dépensées.

L'une des conséquences de l'accent mis sur les résultats a été une initiative pour établir de meilleurs indicateurs. Pour ne citer que deux exemples pertinents : « *Indicators for Monitoring Poverty Reduction* » [Indicateurs pour le suivi de la réduction de la pauvreté] (Carvalho et White, 1994) et le « *Handbook of Democracy and Governance Program Indicators* » [Manuel des indicateurs du programme de démocratie et de gouvernance] (Centre pour la démocratie et la gouvernance, USAID, 1999). Ces initiatives sont louables et devraient être poursuivies car le manque de cohérence des mesures qui persiste dans de nombreux secteurs fait qu'il est difficile de produire des comparaisons pertinentes des résultats obtenus par les programmes.

Plus généralement, dans le domaine du développement international, les « cadres de résultats » sont devenus communs à tous les organismes de développement. [...]

Tout cela était très louable. Il n'y avait qu'un inconvénient : les « résultats » ne mesuraient pas les performances des organismes.

Les mesures de performances peuvent être évaluées en fonction des critères du triple A (alignement, agrégation et attribution[2]). Alignement : les mesures sont-elles conformes aux objectifs de l'organisme? Les mesures des résultats répondent bien à ce critère. Agrégation : les mesures peuvent-elles être agrégées à l'échelle de l'organisme pour donner un chiffre unique de sa performance? Là encore, les mesures des résultats sont efficaces. Attribution : les changements apportés à la

2. Les critères du Triple A ont été proposés dans mon examen de la mesure des performances des organismes de développement présenté dans White (2015a).

mesure peuvent-ils être attribués aux initiatives de cet organisme? Ici, les indicateurs de résultats ne sont pas à la hauteur, comme le montre le cas de l'*USAID*.

En réponse à la GPRA, l'*USAID* a commencé à publier des rapports annuels de performance montrant les résultats en comparaison de leurs objectifs stratégiques, tels que les taux de croissance des principaux bénéficiaires de l'aide étrangère américaine. Dans son examen du rapport de performance de l'année 2000, le *General Accounting Office* (GAO) a écrit à l'*USAID* que les objectifs étaient « si larges et que les progrès étaient affectés par de nombreux facteurs autres que les programmes de l'*USAID*, [que] les indicateurs ne peuvent pas servir de manière réaliste à mesurer les efforts spécifiques de l'agence » (General Accounting Office, 2000). En réponse, l'*USAID* a abandonné l'utilisation d'indicateurs liés aux objectifs stratégiques (« résultats ») pour mesurer la performance de l'*USAID*.

Mon propre engagement à l'égard de ces initiatives est apparu lorsque le *UK National Audit Office* (NAO) m'a demandé d'entreprendre une évaluation du système de mesure de la performance du Department for Internation Development (DFID) en vue de leur propre rapport (National Audit Office, 2002; White, 2002, respectivement), qui concluait que

> l'on peut se demander sur quelles données la direction du DFID fonde ses décisions. Il n'existe pas de système 'ascendant' pour indiquer les performances globales. Et les indicateurs liés aux objectifs du développement international incorporés dans le PSA [le cadre de résultats du DFID] n'ont que peu d'utilité opérationnelle.

Un bref résumé intitulé « *Road to Nowhere* » [La route vers nulle part] avertissait que l'*USAID* avait emprunté la voie des résultats, mais qu'elle était revenue en disant n'y avoir rien trouvé (White, 2005b). Malheureusement, cet appel n'a pas été entendu, et de nombreux organismes et gouvernements de pays en développement ont adopté des cadres de résultats, et certains continuent de le faire.

Mais il y avait aussi autre chose. Le document que j'ai rédigé pour le NAO préconisait l'utilisation de modèles logiques (ou théorie du changement) pour résoudre le problème de l'attribution. Mais il existe un autre moyen : des évaluations d'impact qui mesurent la différence que peut faire une intervention. Il existait déjà quelques études de ce type, mais le nombre d'évaluations d'impact publiées a commencé à augmenter rapidement au cours de la première décennie de ce siècle. L'utilisation de la méthode expérimentale par assignation aléatoire a été particulièrement importante et controversée. C'était la deuxième vague de la révolution de la preuve : la montée des expérimentations par assignation aléatoire.

La deuxième vague : l'augmentation des expérimentations par assignation aléatoire vers 2003

L'utilisation de la méthode expérimentale par assignation aléatoire pour les programmes sociaux n'est pas nouvelle; elle a été utilisée, principalement aux États-Unis, depuis les années 1930 (Oakley, 1998). Mais, à travers le monde, dans tous les secteurs, on observe une nette tendance à la hausse des expérimentations par assignation aléatoire et d'autres modèles d'évaluation d'impact, publié-e-s depuis le début des années 2000.

Dans le domaine du développement international, il y a eu quelques interventions des expérimentations par assignation aléatoire dans les années 1990 – la plus célèbre étant le programme de transfert monétaire conditionnel Progressa au Mexique. Toutefois, le mouvement a pris son essor au début des années 2000. Deux organisations majeures soutenant la méthode expérimentale par assignation aléatoire dans le développement – J-PAL et IPA – ont été fondées en 2003 et 2005 respectivement. Plus notable encore fut l'institutionnalisation de l'évaluation d'impact dans le cadre du *Development Impact Evaluation Program* [Programme d'évaluation de l'impact sur le développement]

DIME, de la Banque Mondiale, en 2004, qui a fourni un financement de pré-amorçage pour soutenir la conception des interventions financées par la Banque Mondiale. Le *Centre for Global Development* (CGD), un think tank basé à Washington, a publié une étude influente *When Will We Ever Learn?* [Quand apprendrons-nous enfin?] qui reprochait à la communauté du développement de dépenser des milliards de dollars dans des programmes pour lesquels il n'existait aucune donnée probante (Levine, Savedoff, et Birdsall, 2006). La campagne du CGD a mobilisé des organismes bilatéraux et des fondations philanthropiques pour appuyer la création de l'*International Initiative for Impact Evaluation* [l'Initiative internationale pour l'évaluation d'impact] (3ie) en 2008. Ces efforts ont entraîné une augmentation substantielle de la production d'évaluations d'impact des interventions de développement international, une tendance qui s'est reflétée dans d'autres secteurs.

En 2008, j'ai quitté la Banque Mondiale et suis devenu le DG fondateur de 3ie. À la Banque mondiale, j'avais dirigé quatre études, mais au cours de mon séjour à 3ie, nous en avons financées près de 200. L'une de nos premières actions a été la création d'une base de données des évaluations d'impact sur le développement. Elle contient désormais près de 5 000 études d'évaluations d'impact. En 2003, moins de 50 évaluations d'impact étaient publiées chaque année, contre plus de 500 en 2012. [...]

Les résultats issus de cette explosion d'évaluations d'impact ont montré l'importance de mener de telles études. Il semble qu'il existe en général une règle de 80 % : 80 % des choses ne fonctionnent pas – c'est-à-dire ont des résultats positifs faibles ou nuls.

Une étude de la Commission européenne a révélé que 85 % des projets financés dans le cadre du Mécanisme de développement propre étaient en réalité peu susceptibles d'entraîner des réductions supplémentaires des émissions de carbone (Cames *et al.*, 2016). L'ONG altruiste basée à Oxford, 80 000 h, a conclu que 80 % était probablement un chiffre généreux – il est plus probable qu'un pourcentage plus élevé de choses ne fonctionnent pas (Todd et The 80,000h team, 2017). Ainsi, en bon-ne-s

Bayésiens et Bayésiennes, en l'absence de preuve du contraire provenant d'une évaluation rigoureuse, nous devrions supposer que notre programme ne fonctionne pas.

Les agences de développement les plus axées sur les données probantes – comme le Département du développement international britannique (DFID) et la Fondation Bill et Melinda Gates – exigent une déclaration de preuves issue d'études rigoureuses pour appuyer les nouvelles propositions et, dans le cas du DFID, la façon dont l'activité proposée permettra de recueillir les preuves nécessaires si elles n'existent pas déjà.

Étant donné que tant de choses ne fonctionnent pas, une évaluation rigoureuse constitue un excellent rapport qualité-prix. L'évaluation du programme mexicain de transfert monétaire conditionnel, Progessa, au milieu des années 90, a coûté 2 millions de dollars. L'évaluation a révélé des effets importants sur l'éducation, la santé, la nutrition et la pauvreté, générant un soutien politique pour le programme qui a ainsi survécu aux transitions politiques. En supposant généreusement que sans l'évaluation, les fonds auraient été utilisés pour un programme deux fois moins efficace, l'utilisation des résultats de l'évaluation a permis à 550 000 enfants supplémentaires de passer à l'école secondaire et à 800 000 enfants âgés de 12 à 36 mois de réduire leur retard de croissance entre 2000 et 2006[3].

Avec tant d'études, il devient difficile de rester au fait des dernières avancées de la littérature. De toute façon, il est peu probable que les décideurs et les décideuses lisent des articles universitaires; mais ils et elles peuvent être influencé-e-s par les résultats d'études très médiatisées. Or, la prise de décision devrait se fonder sur une évaluation de l'ensemble des preuves et non sur des études isolées.

Je prends un exemple, certes controversé, pour illustrer ce point : le déparasitage à l'école.

3. Communication personnelle de Bill Savedoff.

Une étude influente réalisée au Kenya montre les effets importants du déparasitage sur la nutrition, la santé et l'éducation (Miguel et Kremer, 2004). Cette étude en particulier a influencé le mouvement *Deworm the World* [Déparasitons le monde]. Cependant, comme le rapportent les revues systématiques Cochrane (Taylor-Robinson *et al.*, 2015) et Campbell (Welch *et al.*, 2016), la grande majorité des études ne démontrent pas de tels effets. L'exceptionnalisme africain est une énigme. Similairement, la conception qui aiderait à imaginer et cibler les programmes de manière rentable est un casse-tête. Mais pour la plupart des pays du monde, il semble que le déparasitage ne soit pas « le meilleur achat en matière de développement », comme le prétendent certains – nous ne devrions pas être induits en erreur par une seule étude ou par un petit nombre d'études lorsqu'il y a un plus grand nombre de publications.

La troisième vague : l'essor des revues systématiques, 2008

Cette nécessité de s'appuyer sur des éléments de preuve a alimenté la troisième vague de la révolution de la preuve : l'essor des revues systématiques. Dans la plupart des secteurs, cette vague a eu lieu au cours des dix dernières années. Cette vague est arrivée plus tôt dans le domaine de la santé, posant les bases de la médecine fondée sur des preuves, sous l'impulsion de la *Cochrane Collaboration* [Coopération Cochrane] et de l'Organisation Mondiale de la Santé (OMS). D'autres secteurs ont suivi plus récemment.

Une fois encore, cette vague concerne tous les pays et tous les secteurs. Dans le domaine de la politique sociale, peu de revues systématiques ont été publiées avant l'année 2000, environ 25 par an dans les années 90,

puis le chiffre a augmenté dès 2010 pour atteindre 230 en 2016[4]. 3ie a joué un rôle dans l'essor des revues systématiques dans le domaine du développement international. [...]

Certaines revues confirment la vision plutôt pessimiste relative à l'efficacité des programmes. La première étude de Campbell a montré que les programmes *Scared Straight* accroissent en fait la probabilité que les jeunes deviennent des criminels (Petrosino *et al.*, 2013). Un examen des programmes visant à lutter contre les grossesses adolescentes a révélé qu'aucun n'était efficace pour réduire l'activité sexuelle ou les grossesses (Scher, Maynard, et Stagner, 2006).

J'ai quitté 3ie en 2015 pour prendre la direction de Campbell vers la fin de cette année-là. Soutenir la production de revues est l'activité principale de Campbell. Une première étape a été la mise en place d'une nouvelle stratégie comportant deux objectifs clés : davantage de revues et plus d'utilisation de ces dernières.[...]

La difficulté à promouvoir l'utilisation des revues systématiques est dû au fait qu'il s'agit de documents longs et techniques. Ils peuvent également ne pas être accessibles facilement et difficiles à trouver – ou être payants – ou encore difficiles à comprendre. Une revue sur un thème large peut atteindre plusieurs centaines de pages. Et les implications pour les politiques ne sont peut-être pas toujours claires. L'intégration des conclusions des études dans les politiques et les pratiques a constitué la quatrième vague de la révolution de la preuve : le courtage ou la traduction des connaissances.

4. Résultats de la recherche Google Scholar : « revues systématiques » ET sociaux EN Title. Résultats passés au crible jusqu'à cinq pages consécutives sans études éligibles. Recherche effectuée le 12 septembre 2018.

La quatrième vague : l'essor du courtage en connaissances, 2010

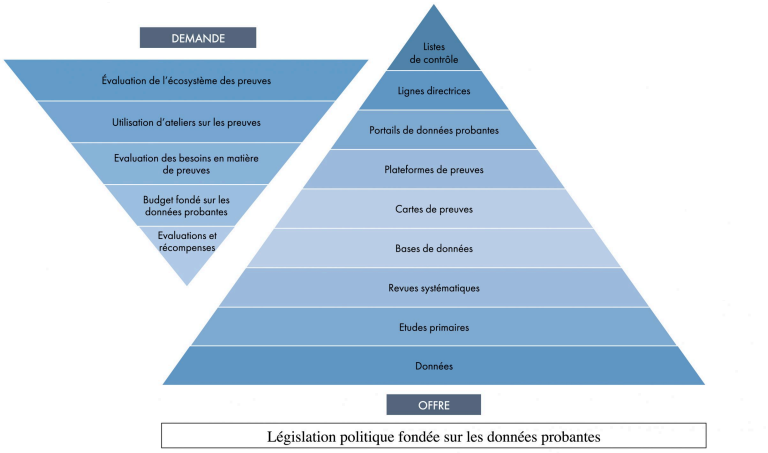
Les activités de la quatrième vague visent à institutionnaliser l'utilisation des données probantes dans les politiques et les pratiques. Il y a deux façons de le faire : l'interaction directe – que j'appelle le modèle nordique – et la création de produits de connaissance comme les portails de données probantes – tel que le mouvement *What Works*. Bien que certaines de ces initiatives soient antérieures à la décennie en cours, c'est cette dernière qui a permis à *What Works* de prendre l'élan nécessaire pour être qualifié de mouvement.

Le Danemark, la Norvège et la Suède disposent chacun de « centres de connaissances » pour l'éducation, la santé et la protection sociale. Il s'agit de centres de recherche financés par le gouvernement. Ce qui est différent dans le modèle nordique, c'est qu'ils disposent d'un personnel dont le travail quotidien consiste à produire des études destinées à éclairer la prise de décision gouvernementale. Il ne s'agit pas de chercheurs et de chercheuses universitaires dont la motivation est de publier. Ce sont plutôt des chercheurs et des chercheuses dont la motivation est de produire des revues systématiques pertinentes pour la politique et la pratique. Les équipes de recherche se réunissent régulièrement avec les organismes gouvernementaux pour convenir des sujets prioritaires et discuter des résultats qui émergent et de la façon dont ils devraient être interprétés à des fins politiques. Ce modèle est également couramment ajusté par des équipes qui fournissent des réponses rapides issues de données probantes plutôt que des revues systématiques complètes dont le nombre augmente.

Le modèle d'interaction directe peut fonctionner lorsque l'on a affaire à un petit nombre de décideurs et de décideuses, par exemple au sein du gouvernement central ou d'un seul organisme. Il est moins adapté

lorsque la prise de décision est décentralisée. Le cas échéant, les produits probants doivent pouvoir être utilisés par les décideurs et les décideuses sans nécessiter aucun soutien.

Figure 2 : L'architecture de la preuve



Mais cette approche se répand. J'y vois la véritable manifestation de la quatrième vague : la construction de l'architecture de la preuve afin d'institutionnaliser l'utilisation de cette dernière. Cette architecture est illustrée à la figure 2. L'institutionnalisation peut être étayée par une législation exigeant que la conception de politiques publiques se fonde sur des données probantes, comme celle adoptée aux États-Unis en décembre 2018 ou la loi mexicaine de 2004 sur le développement social, qui exigeait une évaluation externe de tous les programmes sociaux financés par l'État. Une telle législation exige que les organismes financés par l'État produisent et utilisent des évaluations rigoureuses. [...]

Au Royaume-Uni, le *National Institute of Clinical Excellence and Social Welfare* [Institut national de l'excellence clinique et du bien-être social] (NICE) a recours à des revues systématiques tant pour donner des orientations que pour prendre des décisions sur les dépenses éligibles aux dépenses publiques du *National Health Service* [Service national de santé]. Divers centres britanniques *What Works* ont commencé à produire des lignes directrices, comme celles sur l'utilisation d'assistants et d'assistantes pédagogiques de *l'Education Endowment Foundation* (Sharples, Webster, et Blatchford, s. d.) et les lignes directrices de la police de quartier relatives à ce qui marche pour réduire la criminalité (College of Policing, 2018).

Atul Gawande a plaidé avec éloquence en faveur des listes à cocher dans *The Checklist Manifesto* (Gawande, 2011). Gawande explique comment l'utilisation d'une liste de validation a permis de réduire les « erreurs d'inaptitude » (incapacité à utiliser ce que nous savons) dans tous les domaines, du pilotage des avions à la construction de gratte-ciel. Une telle approche peut-elle fonctionner dans d'autres secteurs? L'expérience des principaux courtiers et courtières en connaissances du mouvement *What Works* suggère que oui.

Le *Teaching and Learning Toolkit* [guide d'enseignement et d'apprentissage] présente des données probantes sur 34 interventions différentes, telles que les cours particuliers. La page d'accueil du guide énumère les 34 interventions avec trois paramètres simples : le coût (indiqué sur une échelle d'un à cinq signes £), la force des éléments probants (indiquée sur une échelle d'un à cinq cadenas) et l'impact. L'impact est indiqué comme le nombre de mois de progrès supplémentaires qu'un-e élève fait lorsqu'il est exposé à cette intervention. Il est de +5 pour les cours particuliers, ce qui signifie qu'un cours particulier a permis, en moyenne, de réaliser des progrès supplémentaires équivalents à cinq mois d'apprentissage. Le meilleur achat consiste à faire aux enfants un retour sur leur travail. Cela coûte très peu et équivaut à 8 mois de progrès supplémentaires. Le « pire

achat » consiste à redoubler une année, ce qui coûte très cher, alors que l'enfant a tendance à faire moins de progrès que s'il n'y avait eu aucune intervention.

Les éléments de preuve présentés dans le guide sont basés sur 34 revues systématiques commandées par l'EEF. Une étude réalisée par le NAO en 2015 a révélé que 64 % – soit près des deux tiers – des écoles utilisaient le guide pour prendre des décisions au sujet des ressources scolaires et des pratiques en classe (National Audit Office, 2015 : 9). En d'autres termes, deux tiers des écoles en Angleterre utilisent des données probantes provenant de revues systématiques pour éclairer leur prise de décision. Tel est le pouvoir d'un courtage en connaissances efficace. […]

Une chose que nous avons apprise des évaluations effectuées dans de nombreux secteurs est que la création de l'offre est rarement suffisante par elle-même – il faut également prêter attention à la demande. La loi de Say selon laquelle l'offre crée sa propre demande ne semble pas s'appliquer dans de nombreux cas, et la promotion de l'utilisation des preuves ne fait pas exception. Ainsi, comme le suggère la figure 2, si nous devons construire le côté offre de l'architecture des preuves, nous devons également tenir compte du côté demande. Ceci est d'autant plus important que les incitations académiques soutiennent l'offre de données probantes issues de la recherche, mais ne récompensent généralement pas les initiatives pour que ces données soient utilisées dans les politiques. […]

Le rôle de l'IA, de l'apprentissage automatique et du Big Data : une cinquième vague?

Les nouvelles technologies offrent un grand potentiel pour étendre la production et l'utilisation de preuves rigoureuses. Le *Big Data* offre des possibilités de collecte de données pour la mesure de l'impact; par

exemple en combinant des données satellitaires et des données pluviométriques pour évaluer les interventions agricoles, ou des données provenant d'appareils de fitness portatifs pour évaluer l'impact des interventions sanitaires ou pour mesurer l'effort de travail des travailleurs ruraux et des travailleuses rurales.

Il est également possible d'améliorer la production de revues systématiques. Des programmes, tels que Rayyan et *EPPI Reviewer*, proposent l'apprentissage automatique pour aider à la sélection d'articles pertinents en vue de leur inclusion dans une revue. *Cochrane Crowd* et *Aidgrade* utilisent le *crowdsourcing* sur le Web pour filtrer et coder des documents avec une méta-analyse automatisée dans le dernier cas. La technologie est déjà disponible pour des revues automatisées toujours à jour, car les algorithmes explorent les bases de données à la recherche d'études pertinentes, mettant à jour les cartes et les revues au fur et à mesure qu'ils les trouvent. L'élément humain peut intervenir lorsque la discrétion ou le jugement d'une experte ou d'un expert est nécessaire, comme dans la production de lignes directrices. Mais le fait que des êtres humains parcourent des articles à la recherche de textes pertinents à inclure est probablement un moyen très inefficace de produire des revues. L'adoption de ces technologies améliorera la rapidité et l'exactitude de la synthèse des preuves.

Il y a aussi des risques. Les machines sont aussi intelligentes que les gens qui sont leur matériau d'apprentissage. Et l'analyse du *Big Data* doit s'appuyer sur une compréhension technique des relations de causalité. Corrélation n'est pas causalité, quelle que soit la taille des données (Elliot *et al.*, 2015). Mais il s'agit là de risques gérables dont les avantages l'emportent sur les risques.

Dernier mot : la preuve est le meilleur moyen d'accéder au développement

La plupart des interventions ne fonctionnent pas, la plupart des interventions ne sont pas évaluées et la plupart des évaluations ne sont pas utilisées. En conséquence, des milliards de dollars provenant de gouvernements et de dons sont gaspillés dans des programmes inefficaces. Financer la recherche sur ce qui fonctionne est le meilleur investissement que nous puissions faire. Joignez-vous à la révolution de la preuve aujourd'hui.

Bibliographie

Boruch, Robert. 2019. in *SAGE research methods foundations*, édité par S. Delamont, P. Atkinson, et A. Cernat. London: Sage Publications.

Cairney, Paul. 2016. *The politics of evidence-based policy making*. London: Palgrave MacMillan.

Cames, Martin *et al.* 2016. « How additional is the clean development mechanism? Analysis of the application of current tools and proposed alternatives ».

Campbell, Donald T. 1969. « Reforms as experiments ». *American Psychologist* 24(4) : 409-29. doi: https://doi.org/10.1037/h0027982.

Campbell, Donald T. 1988. « The experimenting society ». in *Methodology and epistemology for the social science*, édité par E. S. Overman. Chicago: University of Chicago Press.

Carvalho, Soniya, et Howard White. 1994. « Indicators for poverty reduction. World Bank Discussion Paper 254 ».

College of Policing. 2018. « Neighborhood policing guidelines ».

Elliot, Julian H., Jeremy Grimshaw, Russ Altman, Lisa Bero, Steve N. Goodman, David Henry *et al.* 2015. « Making sense of health data ». *Nature* 527(7576) : 31-32. doi : https://doi.org/10.1038/527031a.

Gawande, Atul. 2011. *The checklist manifesto: how to get things right*. London: Profile Books.

General Accounting Office. 2000. « Observations on the US Agency for International Development's Fiscal Year 1999 Performance Report and Fiscal Years 2000 and 2001 Performance Plans ».

Levine, Ruth, William D. Savedoff et Nancy Birdsall. 2006. « When will we ever learn: improving lives through impact evaluation ».

Miguel, Edward, et Michael Kremer. 2004. « Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities ». *Econometrica* 72(1) : 159-217.

National Audit Office. 2002. « Department for international development performance management—helping to reduce world poverty ».

National Audit Office. 2015. « Funding for disadvantaged pupils ».

Oakley, Ann. 1998. « Experimentation and social interventions: a forgotten but important history ». *BMJ* 317(7167) : 1239-42. doi : https://dx.doi.org/10.1136%2Fbmj.317.7167.1239.

Oliver, Kathryn, et Warren Pearce. 2017. « Three lessons from evidence-based medicine and policy : increase transparency, balance inputs and understand power ». *Palgrave Commun* 3(43). doi : https://doi.org/10.1057/s41599-017-0045-9.

Parkhurst, Justin. 2017. *The Politics of Evidence From evidence-based policy to the good governance of evidence*. 1re éd. London: Routledge.

Petrosino, Anthony, Carolyn Turpin-Petrosino, Meghan E. Hollis-Peel et Julia G. Lavenberg. 2013. « Scared straight and other juvenile awareness programs for preventing juvenile delinquency: a systematic review ». *Cochrane Database of Systematic Reviews* 2013(5). doi : 10.1002/14651858.CD002796.pub2.

Scher, Lauren Sue, Rebecca A. Maynard et Matthew Stagner. 2006. « Interventions intended to reduce pregnancy related outcomes among adolescents ». *Cochrane Database of Systematic Reviews* 2(1) : 1-70. doi : https://doi.org/10.4073/csr.2006.12.

Sharples, Jonathan, Rob Webster et Peter Blatchford. s. d. « Making best use of teaching assistants: guidance report ».

Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan et Paul Garner. 2015. « Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance ». *Cochrane Database Syst Rev* 2015(7). doi : 0.1002/14651858.CD000371.pub6.

Todd, Benjamin et The 80,000h team. 2017. « Is it fair to say that most social programmes don't work? » Consulté (https://80000h.org/articles/effective-social-program/).

Welch, Vivian A., Elizabeth Ghogomu, Alomgir Hossain, Shally Awasthi, Zulfi Bhutta *et al.* 2016. « Deworming and adjuvant interventions for improving the developmental health and well-being of children in low and middle-income countries: a systematic review and network metaanalysis ». *Cochrane Database of Systematic Reviews* 12(1) : 1-383. doi : https://doi.org/10.4073/csr.2016.7.

White, Howard. 2002. « A drop in the ocean? The International Development Targets as a basis for performance measurement. Appendix 2 in NAO ».

White, Howard. 2005. « The road to nowhere: results-based management in international cooperation ». in *Why did the chicken cost the road? And other stories on development evaluation*, édité par S. Cummings. Amsterdam: KIT Publishers.

2. La logique de l'évaluation

MICHAEL SCRIVEN

[Traduit de : « Logic of Evaluation », in Scriven, Michael. 1991. *Evaluation thesaurus* (p.216-223). Sage. Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

La fonction clé de l'inférence évaluative est de parvenir valablement à des conclusions évaluatives à partir de prémisses factuelles (et d'éléments de définition bien sûr); la principale tâche de la logique de l'évaluation est donc de montrer comment cela peut se justifier. Cette tâche a été et est toujours considérée comme impossible par la plupart des logicien-ne-s et des scientifiques – en particulier les spécialistes des sciences sociales. La première partie de cet article aborde le problème d'un point de vue pratique, en exposant deux paradigmes largement utilisés et respectables. La deuxième partie examine certains problèmes partiellement techniques liés à l'extension des paradigmes à d'autres domaines d'évaluation, et la troisième partie aborde le problème dans le langage technique du/de la logicien-ne et du/de la philosophe des sciences. Enfin, une référence est faite à un ou deux autres sujets de la logique de l'évaluation.

1. Quels que soient les mérites de la discussion entre les logicien-ne-s, les évaluations quotidiennes de produits démontrent la faisabilité de l'inférence des faits aux valeurs. Elles partent de faits concernant les performances de divers produits et tirent des conclusions sur leur mérite relatif ou absolu. On peut difficilement prétendre que chaque numéro de *Consumer Reports*[1] soit rempli de mensonges.

1. NdT : Revue testant des produits du point de vue des consommateurs.

Pour un usage pratique, le paradigme de l'évaluation de produits est solide et généralisable. Si l'on en doute, on peut plutôt se tourner vers l'équivalent de l'évaluation de produits pour le scientifique en activité : l'évaluation de données, de conceptions, d'hypothèses, d'instruments, d'articles soumis pour publication, etc. Dans chaque cas, la ou le scientifique travaille à partir de données factuelles sur les performances et arrive à une conclusion évaluative; si elle ou il est contesté-e, il ou elle n'a aucun problème à défendre sa conclusion en faisant appel à des preuves, des définitions et des déductions valides.

2. La critique habituelle faite au paradigme de l'évaluation de produits en tant qu'exemple de la manière d'arriver à des conclusions évaluatives à partir de prémisses factuelles suggère qu'il repose sur des valeurs partagées par ses lecteurs, ce qui n'est pas transposable à l'évaluation des programmes par exemple. Les gens ne sont pas radicalement en désaccord sur ce qu'ils ou elles apprécient dans un détergent; mais ils ou elles ne sont pas d'accord sur ce qu'ils ou elles veulent d'une clinique pour toxicomanes, d'une patrouille de police ou d'un programme scolaire. Cette critique comporte deux erreurs. En premier lieu, ce ne sont pas les valeurs partagées qui appuient la validité des évaluations de produits. Vous aurez remarqué que les associations de consommateurs et de consommatrices réalisent rarement, voire jamais, d'enquête pour vérifier ce que les gens apprécient dans les appareils et les produits. Les scientifiques ne font pas non plus d'enquêtes sur ce qui fait une bonne théorie. Ce n'est pas parce qu'ils ou elles pensent avoir des intuitions infaillibles sur les préférences de leurs pairs. C'est parce qu'ils ou elles partagent une même compréhension de la signification des termes décrivant le sujet évalué. Si vous savez ce qu'est une montre, vous savez que la précision de la mesure du temps, la lisibilité et la durabilité sont des qualités essentielles; et si vous savez cela, vous savez comment établir de prime abord certaines conclusions évaluatives à partir de prémisses factuelles sur le mérite comparatif des montres. (Il en va de même dans le cas des théories scientifiques.) Les noms de la plupart des produits et des entités méthodologiques ont une logique quelque

peu similaire à celle des idéaux-types– des entités bien connues du domaine scientifique où le « gaz idéal », le « ressort parfaitement élastique », le « col bleu », la « concurrence parfaite », etc., servent un but utile. Ainsi, les conclusions relatives aux montres, aux lave-vaisselles, aux théories, etc., découlent directement de la compréhension de la signification des termes (leur définition implicite, les idéaux intégrés à la compréhension du concept) et des faits relatifs à leurs performances.

Ce n'est pas parce que certaines consommatrices ou certains consommateurs font preuve de goûts « aberrants » que ces conclusions sont erronées – pensez à ces personnes qui achètent des montres Rolex *DayDate* à 20 000 $, malgré le fait qu'elles sont beaucoup moins précises, plus difficiles à lire, nécessitent un entretien plus fréquent et plus coûteux, et sont plus susceptibles d'inciter à des agressions physiques sur le porteur que la Microtec suisse à 80 $ (l'actuelle championne de la précision et de la luminosité). L'existence de personnes cherchant à acheter un statut ne montre pas qu'il est erroné de qualifier la Microtec de meilleur achat, ni même qu'il est erroné de la classer au sommet des montres en termes de mérite. La légitimation de ce type d'inférence fait partie de la logique des descripteurs des produits, qui sont des abstractions issues d'un ensemble d'indicateurs incluant différentes fonctions. Un autre élément de la même logique est le regroupement par prix : nous protégerons souvent l'évaluation d'un produit contre les attaques en introduisant des catégories de prix. Nous pouvons alors reconnaître la supériorité à un prix plus élevé, comme lorsque nous disons que la Lexus 400 est une meilleure voiture que la Nissan Maxima, mais nous nous réservons le droit de dire que la Maxima est la meilleure voiture dans sa catégorie de prix. La Rolex, quant à elle, n'est pas une meilleure montre que la Microtec. Elle n'est qu'un meilleur symbole de statut social – et ce uniquement parmi les personnes ayant une appréciation limitée de la technologie – ce qui n'a pratiquement rien à voir avec son mérite en tant que montre.

Les études de performance en situation réelle, principales sources de faits dans l'évaluation de produits, impliquent plus que la vérification des performances sur les critères intégrés dans le sens commun. Les tests simulés ou réels sur le terrain (i) révèlent d'autres critères qui sont de toute évidence pertinents (généralement confirmés à l'unanimité par le personnel du laboratoire, mais la tenue d'un atelier peut être envisagée) et (ii) permettent d'établir un lien entre les critères initiaux et les nouveaux critères avec les mesures empiriques. Par exemple, bien qu'elle ne fasse (peut-être) pas partie de la signification spécifique d'une « ampoule électrique », la sécurité électrique est certainement un critère de mérite pour ces ampoules, et des mesures de ce type doivent être élaborées. (iii) Les essais contribuent également à la procédure de « découpage en tranches de prix », c'est-à-dire à l'identification de bons points de coupe pour les catégories de coûts (par exemple, voitures économiques, voitures de luxe) et au découpage en sous-catégories définies par fonction telles que « berline familiale », « fourgonnette », « voiture de sport », etc. L'introduction de sous-catégories préserve les conclusions évaluatives de l'accusation d'invalidité en leur substituant une validité limitée. Toutefois, il ne s'agit là que de raffinements; l'essentiel est que le paradigme de l'évaluation de produits survit à l'attaque de l'inférence faits-valeur en utilisant l'analyse fonctionnelle, plutôt que les faits variables et insaisissables que pourrait apporter une enquête portant sur les valeurs des consommateurs, pour établir ses conclusions évaluatives à partir de prémisses factuelles. Notre langage définit implicitement les idéaux-types dans le domaine des produits, comme il le fait souvent dans le domaine psychologique et sociologique, et nous les utilisons, avec le type de raffinement indiqué, comme les normes en fonction desquelles nous évaluons les produits réels. Les idéaux-types eux-mêmes sont basés sur une analyse fonctionnelle et définitionnelle, et non sur des sondages de popularité. Le même modèle que nous utilisons dans l'évaluation de produits s'applique – avec des modifications mineures – à l'évaluation des candidatures en réponse à une offre de poste, aux plans de construction par le biais de spécifications et aux programmes sociaux de la même manière (voir ci-dessous).

Ainsi, l'évaluation ne se cache pas dans des prémisses douteuses ni dans des hypothèses arbitraires sur ce qui est bon et mauvais, et encore moins sur ce qui est considéré comme bon et mauvais. Il suffit d'utiliser les « définitions » habituelles, c'est-à-dire les conceptions d'entités fonctionnelles, dont une partie de la conception est qu'elles sont de meilleurs exemples de leur genre si elles remplissent mieux les fonctions qui les définissent, ce qui en est en soi une vérité définitionnelle.

Mais qu'en est-il des considérations éthiques? Faut-il dénigrer les produits dont les contenants n'utilisent pas de matériaux recyclés? Ceux qui pourraient blesser des enfants curieux, bien qu'il n'y en ait pas dans notre famille? On pourra répondre gentiment que le rôle de l'éthique ici n'est pas différent de son rôle dans toute activité professionnelle; elle a un rôle et des codes professionnels existent – ou devraient être créés – pour le préciser. La réponse abrupte consiste à dire que les considérations éthiques ne sont que des considérations générales de stratégie sociale (analogues à des considérations juridiques) et que les stratégies sociales font l'objet d'une évaluation comme toute politique [...]. Ainsi, dans la mesure où l'éthique entre en jeu, les questions éthiques doivent être réglées avant que la tâche ne soit achevée; et leur résolution est aussi un problème d'évaluation. Cela n'est pas différent du fait que les questions relatives au personnel ou aux questions fiscales ou juridiques doivent être réglées avant que nous puissions tirer des conclusions évaluatives finales au sujet d'un programme ou d'une institution – ou d'une guerre.

Un autre problème auquel il faut réfléchir concerne la relative imprécision du concept de la « fonction correcte d'une clinique de traitement de la toxicomanie » par opposition à la « fonction correcte d'un stylo à bille ». Une bonne analogie ici est avec la « fonction appropriée du MMPI[2] (ou de tout autre test standardisé) ». Cette fonction ne se limite pas à la fonction initialement prévue ou modifiée de ce test (l'erreur habituelle d'une évaluation qui viserait à vérifier l'atteinte des objectifs),

2. NdT : Il s'agit d'un test standardisé de la personnalité.

mais est fonction d'une interaction entre les besoins existants et les ressources disponibles. Fondamentalement, lorsque nous procédons à l'évaluation d'un programme, nous devons déterminer simultanément la meilleure fonction et le mérite du programme. Ce n'est pas un processus banal, mais ce n'est pas plus problématique que de faire la même chose avec un test psychologique ou un instrument scientifique.

Mais supposons que nous devions nous rabattre sur des enquêtes sur les valeurs. Même dans ce cas, il est possible de faire preuve d'une plus grande objectivité. Supposons, par exemple, que l'on réalise une enquête sur les préférences dans le cadre d'une évaluation des besoins, ou dans le cas où les désirs sont les paramètres moteurs des choix (c.-à-d. lorsque l'éthique n'intervient pas). Supposons qu'il s'avère que les personnes interrogées ont un large éventail d'opinions différentes sur ce qui est souhaitable. Supposons, en outre, que les performances des candidates et des candidats selon des critères différents ne soient pas toutes identiques. Il s'agit d'une situation assez courante et souvent évoquée comme une raison de penser qu'il ne peut y avoir d'objectivité dans les évaluations : « le meilleur X sera très différent selon les personnes ». En fait, même sans recourir aux procédures habituelles de ségrégation, de sélection et d'idéalisation, les résultats sont souvent extrêmement solides et généralisables. En d'autres termes, le meilleur X sera le meilleur pour tou-te-s les répondant-e-s. Cela se produit évidemment lorsqu'un-e candidat-e surpasse les autres sur tous les critères, puisque les différences de pondération des critères ne sont alors plus pertinentes. Mais cela se produit également dans un très grand nombre de cas lorsque plusieurs candidats gagnent sur l'un ou l'autre critère, mais où leur avance sur ces critères est telle que, même multipliée par les (différentes) pondérations du critère, elle ne suffit pas à compenser l'avance du/de la candidat-e principal-e. Par conséquent, aucune conclusion relativiste au sujet de l'évaluation ne découle du fait qu'il existe de grandes différences dans les valeurs des consommateurs/trices, qu'elles soient ou non associées à de grandes différences dans les performances des candidat-e-s sur les dimensions valorisées. Il peut toujours y avoir – et c'est souvent

le cas – des gagnants ou gagnantes absolu-e-s, dont on peut dire qu'ils ou elles sont les meilleur-e-s pour tout le monde. Il s'agit de cas où les gagnant-e-s écrasent tout simplement l'opposition.

Trois questions techniques. (i) Il est clair qu'un rôle central est joué ici par la notion de concepts groupés ou de « définitions par critères », contrairement aux définitions classiques qui étaient des règles de substitution ou des ensembles de conditions logiquement nécessaires et suffisantes. Par exemple, on dit que le sens du mot « montre » s'appuie sur des critères définitionnels comme la capacité de rester à l'heure. La plupart des termes utilisés dans la langue commune et dans les langues techniques des disciplines sont des concepts groupés. Ce fait réduit à néant le soi-disant « argument de la question ouverte » de G.E. Moore, qui était censé montrer que la signification des termes évaluatifs ne pouvait pas être « réduite » à des concepts non évaluatifs. Les réductions suggérées étaient censées commettre « l'erreur naturaliste », mais ne le faisaient que si elles étaient si simplistes qu'elles ne valaient pas la peine d'être examinées. (Référence « The Logic of Criteria », *Journal of Philosophy*, octobre 1959, réimprimé dans Critères, éd. John V. Canfield [Garland, 1986]). (ii) La reconnaissance de la nature et de l'ubiquité des concepts groupés conduit également à la notion **d'inférence probante**, le concept plus général d'inférence qui englobe l'inférence inductive et évaluative. (iii) L'inférence probante génère des conclusions à première vue plutôt que des conclusions catégoriques, conditionnelles ou (quantitatives) probabilistes. L'inférence probante peut être utilisée pour générer des conclusions en utilisant la signification fondamentalement *qualitative* de la probabilité (« c'est une pomme, donc l'intérieur est probablement d'une couleur très différente de la peau »), dont les versions plus mathématiques dérivent dans des cas particuliers, et elle est donc liée à un type d'inférence inductive. L'inférence à la meilleure explication est également une inférence probante, de même que la plupart des inférences à des conclusions juridiques ou évaluatives. (iv) L'un des aspects de l'inférence probante est sa nature itérative ou potentiellement itérative. C'est-à-dire qu'une première série d'inférences

probantes génère des conclusions à première vue, qui sont testées par une enquête plus approfondie et modifiées à la lumière de nouvelles données, atteignant progressivement des niveaux de confiance justifiés, sans jamais dépasser la possibilité d'erreurs empiriques. Ce trait, si caractéristique du processus de raisonnement juridique, est tout aussi caractéristique de l'inférence évaluative par ses longues listes de contrôle multidimensionnelles. Elle est également caractéristique d'une grande partie du raisonnement scientifique, bien que les scientifiques semblent l'oublier lorsqu'ils évoquent la nature *prima facie* des conclusions évaluatives comme un signe que l'inférence y amenant n'est pas réellement scientifique. On entend souvent la question suivante : « Comment savez-vous qu'il n'y a pas d'autres considérations qui l'emporteront sur celles-ci? » Réponse : pour la même raison que vous connaissez parfois l'explication d'un phénomène physique. Vous recherchez des alternatives; même dans ce cas, vous ne pouvez jamais être absolument sûr-e, mais vous pouvez devenir de plus en plus sûr-e par une enquête itérative minutieuse, tout comme dans le processus de confirmation d'une hypothèse provisoire dans une enquête scientifique (ou criminelle).

Compte tenu de tout cela, que pouvons-nous dire, dans le cadre technique, de l'inférence depuis des prémisses factuelles jusqu'à des conclusions évaluatives? Il semble évident que cela ne peut pas se faire par une déduction stricte, mais en réalité presque aucune inférence scientifique ou de bon sens n'est déductive. Si l'on accepte l'idée que le seul autre choix est l'induction, et que l'on est impressionné-e par l'affirmation de Popper selon laquelle il n'y a pas de logique d'induction – seulement des suppositions et des confirmations– cela met un terme à la discussion. Il ne semble y avoir que trois options possibles. (A) On peut trouver un moyen de contourner les arguments de Popper et établir que l'inférence évaluative est inductive; (B) On peut inventer un nouveau type de logique, ce qui risque de ne faire que contourner la question (pourquoi devrait-on supposer que donner un nouveau nom à une erreur la rende légitime?); Ou (C) on peut essayer de sortir de sa manche un tour

de magie déductif qui a semblé logiquement impossible aux meilleur-e-s logicien-e-s des deux derniers siècles. En fait, on peut faire les trois en toute légitimité.

(A) Popper a certainement tort à propos de la logique de l'induction – ironiquement, il était encore sous le charme du paradigme déductif. Il y a une logique d'induction, bien que ses principes ne puissent être formulés de la même manière que ceux de la logique déductive. On peut s'y former, l'enseigner et l'évaluer, et tou-te-s les scientifiques peuvent à chaque instant l'exécuter avec compétence, certain-e-s même avec brio (à cet égard, elle n'est pas différente de la déduction). Ses normes sont celles de l'argument scientifique; ses concepts de base sont dirigés par le concept d'explication – l'inverse du soutien inductif – et son assistant principal, le concept de définition critérielle, l'inverse de l'inférence de prime abord. Il s'agit pour l'essentiel d'une logique implicite – tout comme la grammaire d'une langue est pour la plupart implicite, mais assez précise pour que nous puissions créer et distinguer les phrases grammaticalement des phrases incorrectes dans presque tous les cas. Les outils de l'argument inductif et de la critique sont des analogies, des exemples, des contre-exemples, des contre-explications et des contrastes, plus souvent que des règles exactes, et les déclarations qu'elle utilise – comme les « règles de grammaire » – ne sont que des guides approximatifs de la vérité, c'est-à-dire des allusions et des heuristiques plutôt que des lois exactes. Nous utilisons certains paramètres comme « *prima facie* » (de prime abord), l'équilibre entre les éléments de preuve, et « toutes choses étant égales par ailleurs » – parfois « probablement » pour signaler les qualifications impliquées. L'un des exemples paradigmatiques du raisonnement inductif est le raisonnement évaluatif, et il suffit de consulter les *Consumer Reports* pour voir comment il fonctionne.

(B) Il est peut-être plus sain de commencer plus en arrière, plus près des fondamentaux, et de formuler tout cela sous l'angle d'une nouvelle logique qui couvre une grande partie de notre raisonnement quotidien ainsi que le raisonnement scientifique et juridique. Dans l'une de ces approches – la logique probante – la logique est traitée comme étant nécessairement

et essentiellement une grammaire, avec parfois des cas limitatifs simples – les « règles grammaticales » occasionnelles, d'une part, et les règles de la logique déductive, d'autre part. (Les mathématiques sont, de ce point de vue, un pas au-delà de la logique déductive en direction de la science, mais pas aussi loin que Mill l'avait supposé). Dans la logique probante, le contexte est aussi important que le contenu; dans la logique traditionnelle, la nature de la logique est d'être indépendante du contexte. Dans la logique probante, les définitions ne sont jamais des règles de substitution, mais seulement des explications du sens, susceptibles d'être indéfiniment reformulées et affinées par ceux qui comprennent les termes définis, chaque fois qu'elles ne parviennent pas à transmettre le sens. À ce compte, le raisonnement évaluatif est un raisonnement probant typique, comme la plupart des inférences juridiques, de bon sens et scientifiques. (On trouvera un compte-rendu étendu, bien qu'encore programmatique, de ce phénomène dans « *Probative Logic* » [Logique Probante], dans *Argumentation : Cross the Lines of Discipline*, édité par van Eemeren, Grootendorst, Blair et Willard [Foris, 1987]).

(C) Enfin, on peut affirmer (comme le fait John Searle) qu'il y a des cas, quoique rares, où la déduction directe peut être utilisée pour briser le tabou. (i) Le meurtre est défini comme une mise à mort injustifiée (probablement plus proche de l'usage correct que la définition habituelle du dictionnaire qui le définit comme une mise à mort illégale). « Injustifiée » signifie ici en gros « pas en cas de légitime défense, de guerre, d'exécution, ou pour sauver la vie d'autrui ». (ii) On peut parfois établir, peut-être avec l'aide d'aveux, qu'un meurtre pour des raisons égoïstes, commis par quelqu'un-e qui n'est pas dans une situation désespérée, s'est produit. (iii) Nous pouvons donc conclure, à partir des définitions et des faits, que l'agent ou l'agente responsable est un meurtrier ou une meutrière, une conclusion évaluative. Il existe des moyens ultimes de contester cet exemple, notamment en s'attaquant à la définition employée en contexte (comme dans le cas de la définition de « injustifiée »), mais il s'agit bien de moyens ultimes puisqu'ils impliquent

l'abandon d'une grande partie de la pratique du dictionnaire afin de sauver un dogme logique. Ce cas est étroitement analogue à l'inférence standard dans l'évaluation des produits.

Il devrait donc être clair que la logique d'une évaluation pratique sérieuse n'est pas l'inférence déductive invalide de « j'aime X » à « je devrais avoir X » (ou « je devrais obtenir X » ou « je mérite X »).

Il est vrai que, dans un contexte approprié, ce cas simple constitue le cas limite d'un type d'inférence *prima facie*, assez courant dans l'évaluation de produits, à savoir l'inférence des attributs que l'on souhaite avoir dans un produit à la conclusion que l'on devrait acheter un produit particulier. Il est juste de dire qu'il y a de nombreux pièges possibles sur le chemin qui mène de cette prémisse à cette conclusion, et la logique de l'évaluation est consacrée à gérer ces pièges.

Plusieurs autres questions relèvent de la logique de l'évaluation, comme la nature des évaluations des besoins – ces derniers semblent souvent relever de prémisses de valeur mais semblent aussi être des questions factuelles – et le problème de la spécification de l'objet logique parfois très complexe que l'on décrit comme une évaluation – la question des **paramètres d'évaluation**. (Certaines de ces questions ont fait l'objet d'une discussion plus approfondie dans *The Logic of Evaluation*, Edgepress, 1981.) [...]

Bibliographie

Michael Scriven. 1987. « Probative Logic ». in *Argumentation: Across the Lines of Discipline*, édité par F. H. van Eemeren, R. Grootendorst, J. A. Blair et C. A. Willard. Providence: Foris Pubns USA.

Scriven, Michael. 1981. *The Logic of Evaluation*. 2e éd. Inverness: Edgepress.

Scriven, Michael. 1986. « The Logic of Criteria ». in *Criteria*, édité par J. V. Canfield. New York: Garland.

3. Liste des valeurs et critères de l'évaluation

DANIEL L. STUFFLEBEAM

[Traduit de Stufflebeam, Daniel L. 2001. « Evaluation Values and Criteria Checklist », Western Michigan University Evaluation Checklists (ressource en ligne). Traduction par Carine Gazier et Thomas Delahais. Article originellement publié en *open access*.]

Les bonnes évaluations sont fondées sur des **valeurs** (principes, attributs ou qualités considéré-e-s comme intrinsèquement bon-ne-s, désirables, important-e-s et ayant du mérite de façon générale) et des **critères** (normes sur lesquelles fonder les jugements) clairs et appropriés. Cette liste de contrôle est destinée à aider les évaluateurs/trices, et leurs client-e-s à considérer une gamme pertinente de valeurs et de critères génériques lorsqu'ils ou elles identifient ceux qui sous-tendront des évaluations particulières.

Valeurs sociétales

Équité : Juste [*Fair*] pour tou-te-s – une conformité libre et raisonnable aux normes acceptées du droit naturel, de la loi et de la justice, sans préjugé, favoritisme ou fraude, et sans imposer de contraintes excessives en matière d'accès.

Efficacité : Réussit à répondre aux besoins ciblés et à atteindre les objectifs.

Conservation : Des efforts délibérés, réfléchis et fructueux pour éviter le gaspillage et préserver les ressources naturelles et économiques, afin que les institutions/programmes puissent fonctionner avec un bon rapport coût-efficacité et que les villes et les campagnes puissent convenir aux générations futures.

Excellence : Viser des normes exigeantes et réaliser des performances proches de celles-ci ou posséder un nombre remarquable de bonnes qualités.

Citoyenneté : Être un élément constructif, agir de manière responsable et contribuer au bien-être commun de sa communauté.

Liberté : Les droits inaliénables des citoyen-ne-s à suivre leur conscience en utilisant, soutenant et agissant selon leurs croyances dans des limites raisonnablement formulées et légalement spécifiées et sans contraintes excessives.

Légalité : Respecter les lois en se comportant bien, en réglant les différends, en distribuant les biens publics, en maintenant l'ordre et en sanctionnant ou punissant les mauvais comportements.

Défense nationale : Maintenir la capacité à protéger la société et les citoyen-ne-s contre les agressions extérieures et les atteintes intérieures, afin de protéger les valeurs et les possessions de la société, mais aussi, les droits internationaux, ainsi qu'un statut viable dans la communauté mondiale; et de préserver la liberté et les autres droits des citoyen-ne-s.

Critères inhérents à la définition de l'évaluation

Mérite : La valeur ou la qualité intrinsèque d'un objet; il s'agit de savoir si un programme, un produit ou un service correspond à l'état de l'art en matière de concept, de modèle, de livraison, de matériaux et de résultats.

Intérêt [*worth*] : La valeur extrinsèque d'un objet ou la mesure dans laquelle il est utile et abordable pour répondre aux besoins évalués d'un groupe défini de bénéficiaires. Alors que toutes les institutions devraient s'efforcer d'offrir des services méritoires, elles doivent parfois mettre fin à des programmes même bons ou à d'excellent-e-s membres du personnel, parce que les électeurs/trices de l'institution n'ont pas besoin de leurs services ou ne peuvent pas se les offrir.

Critères inhérents au modèle d'évaluation du CIPP

[Le CIPP est un modèle qui appelle à l'évaluation du contexte, des intrants, du processus et du produit dans le processus de jugement de la valeur d'un programme[1].]

Objectif défendable : Un objectif qui est éthique, socialement responsable, réalisable et bénéfique pour la société ou les individus.

Besoins : Conditions ou choses qui sont nécessaires ou utiles pour atteindre un objectif défendable, par exemple, la capacité d'un-e enfant à lire et la présence d'enseignant-e-s compétent-e-s dans une école.

Plan justifiable : Un ensemble de dispositions solides, ciblées et réalisables pour atteindre un objectif défendable; doit répondre aux besoins des bénéficiaires.

Mise en œuvre responsable : Cohérence entre les activités et les plans et entre les dépenses et le budget, y compris l'amélioration des plans et des budgets si nécessaire.

1. Voir Stufflebeam, D. L. 2020. "The CIPP Model for evaluation", dans D. L. Stufflebeam, G. F. Madaus et T. Kellaghan (éds.). *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 279-317). Boston: Kluwer Academic Publishers.

Résultats louables : Des résultats de haute qualité, un service à tous les bénéficiaires légitimes, importance, sécurité et bon rapport coût-efficacité.

Valeurs institutionnelles

Mission : La fonction principale d'une organisation ou d'une institution.

Finalités : Les résultats souhaités, généralement sur le long terme, vers lesquels l'ambition et l'effort sont dirigés.

Priorités : Notations préférentielles attribuant l'attention, le temps et les ressources à des programmes, des objectifs ou d'autres entités avant les alternatives concurrentes.

Exigences techniques

Codes : Ensembles de règles de procédures et de normes de matériaux conçus pour assurer l'uniformité et protéger l'intérêt public dans des domaines tels que la construction de bâtiments et la santé publique, généralement établis par un organisme public et ayant généralement force de loi dans une juridiction particulière.

Normes : Ensembles de principes, de règles ou d'attentes en matière de comportement ou de pratiques professionnelles – dans des domaines tels que les spécialités médicales, le droit, l'ingénierie, les tests éducatifs et les écoles primaires et secondaires – établis par des groupes organisés et parfois renforcés par certains pouvoirs de sanction des groupes à l'encontre des membres non conformes.

Responsabilités du personnel

Compétence professionnelle : Les obligations d'un individu associées à l'appartenance à une profession (par exemple, on attend des enseignant-e-s qu'ils ou elles maintiennent à jour leurs connaissances dans leurs domaines de compétence, qu'ils ou elles développent leur capacité à gérer des classes, qu'ils ou elles soient compétent-e-s pour mesurer les résultats scolaires, qu'ils ou elles soient habiles pour communiquer avec les élèves et les parents, qu'ils ou elles démontrent leur efficacité à aider les élèves à apprendre et qu'ils ou elles contribuent à faire progresser l'enseignement en tant que profession).

Performance professionnelle : L'accomplissement des responsabilités professionnelles assignées (par exemple, on peut attendre d'un-e enseignant-e qu'il ou elle enseigne efficacement les cours assignés, maintienne les affichages dans la classe, gère les activités parascolaires, conseille les étudiant-e-s, communique avec les parents et coopère aux projets d'amélioration de l'école).

Critères de terrain

Critères idiosyncrasiques : Ne peuvent être spécifiés à l'avance, doivent être négociés, doivent être définis de manière très détaillée sur le plan opérationnel (par exemple, l'évaluation d'un programme de vulgarisation agricole dans une certaine localité doit non seulement évaluer ses mérites par rapport aux normes de la technologie pédagogique, mais aussi déterminer avec les agriculteurs/trices de la région les critères d'évaluation qu'ils ou elles apprécient, par exemple, dans quelle mesure le programme agrémente plutôt qu'il ne duplique d'autres sources d'information qu'ils ou elles obtiennent et surtout dans quelle mesure il répond à leurs besoins d'information les plus importants).

4. Lignes directrices et repères pour une évaluation constructiviste

EGON G. GUBA ET YVONNA S. LINCOLN

[Traduit de : Guba, Egon G. et Yvonna S. Lincoln. 2001. « Guidelines and checklist for constructivist (a.ka. fourth generation) evaluation », Western Michigan University Evaluation Checklists (ressource en ligne). Traduction par Carine Gazier et Thomas Delahais. Article originellement publié en *open access*.]

[Remarque : Les lignes directrices et les listes de contrôle pour les évaluations constructivistes et les rapports, présentées ci-après sont basées sur Egon G. Guba et Yvonna S. Lincoln. 1989. *Fourth Generation Evaluation* [Evaluation de quatrième génération], Newbury Park, CA: Sage Publications. Des informations générales utiles peuvent être trouvées dans Yvonna S. Lincoln et Egon G. Guba. 1985. *Naturalistic Inquiry* [Enquête naturaliste] Beverly Hills, CA: Sage Publications.]

Définition de l'évaluation

L'évaluation est l'une des trois formes fondamentales d'enquête systématique, les autres étant la recherche et l'analyse des politiques publiques. C'est une forme d'enquête dont l'objet est un evaluanda[1] (un programme, un processus, une organisation, une personne, etc.) et qui aboutit à des constructions (jugements) portant sur son « mérite » [merit] ou son « intérêt » [worth]. Les constructions relatives au mérite se

1. NdT : Evaluanda (evaland) est un mot générique désignant tout objet de l'évaluation.

concentrent sur la qualité intrinsèque d'un evaluanda, indépendamment de l'environnement dans lequel il peut trouver des applications. Les constructions relatives à l'intérêt portent sur l'utilité ou l'applicabilité extrinsèque d'un evaluanda dans un contexte local concret. L'évaluation d'un evaluanda encore au stade de la proposition initiale ou en cours de développement est dite « formative », tandis que l'évaluation d'un evaluanda ayant atteint la maturité est dite « sommative ».

Définition de l'évaluation constructiviste

L'évaluation constructiviste est une forme d'évaluation basée sur les propositions (hypothèses de base) qui sous-tendent le paradigme constructiviste. Le paradigme constructiviste diffère des autres paradigmes de la connaissance couramment utilisés, y compris les paradigmes scientifiques, artistiques, religieux, juridiques et d'autres paradigmes similaires. Il repose sur trois hypothèses fondamentales, communément appelées ontologique, épistémologique et méthodologique, à savoir :

- L'hypothèse ontologique de base du constructivisme est le relativisme, c'est-à-dire que la compréhension humaine (sémiotique) qui organise l'expérience pour la rendre apparemment compréhensible et explicable, est un acte d'interprétation et est indépendante de toute réalité fondamentale. Dans le cadre du relativisme, il ne peut y avoir de vérité « objective ». Cette observation ne doit pas être considérée comme une position « tout est permis »; voir la section sur les critères ci-dessous.
- L'hypothèse épistémologique de base du constructivisme est le subjectivisme transactionnel, c'est-à-dire que les affirmations sur la « réalité » et la « vérité » dépendent uniquement des ensembles de signification (informations) et du degré de sophistication dont disposent les individus et les publics engagés dans la formation de ces

affirmations.

- L'hypothèse méthodologique de base du constructivisme est l'herméneutique-dialectique, c'est-à-dire un processus par lequel les constructions entretenues par les différents individus et groupes impliqués (parties prenantes) sont d'abord découvertes et sondées quant à leur signification, puis confrontées, comparées et contrastées dans des situations de rencontre. Le premier de ces processus est l'herméneutique; le second est la dialectique. Voir les sections « découverte » et « assimilation » ci-dessous. Notez que cette hypothèse méthodologique est silencieuse sur le sujet des méthodes et, en particulier, sur le sujet des méthodes « quantitatives » vs « qualitatives ». Les deux types de méthodes peuvent être et sont souvent pertinents pour toutes les formes d'enquêtes évaluatives.

Il n'est pas approprié de combiner les paradigmes dans la conduite d'une évaluation, par exemple en utilisant à la fois des propositions scientifiques (positivistes) et des propositions constructivistes dans le cadre d'une même étude. Il ne s'agit pas d'un appel à la « pureté » ni d'une volonté d'exclusion. Il s'agit simplement d'une mise en garde contre le fait que le mélange de paradigmes peut aboutir à des approches et des conclusions absurdes.

Les deux phases de l'évaluation constructiviste : découverte et assimilation

La phase de découverte de l'évaluation constructiviste représente l'effort de l'évaluateur pour décrire « ce qui se passe ici », « ici » étant l'évaluanda et son contexte. La phase de découverte peut ne pas être nécessaire (ou n'être nécessaire qu'à minima) s'il existe une ou plusieurs constructions préexistantes relatives au sujet évalué sur lesquelles on peut s'appuyer (par exemple, à partir d'une évaluation antérieure ou d'une proposition de projet), c'est-à-dire que certaines significations (informations) et un

certain degré de sophistication dans leur interprétation sont déjà disponibles. Il existe de nombreuses façons de répondre à la question de la découverte, selon les constructions pertinentes et préexistantes que l'évaluateur et les informateurs et répondants locaux apportent à l'enquête. Les découvertes sont elles-mêmes des organisations sémiotiques, c'est-à-dire des constructions mentales. ATTENTION : Si les constructions préexistantes proviennent de sources extérieures à l'évaluation du sujet, et en particulier de la littérature professionnelle, il convient d'évaluer leurs bases paradigmatiques; si ces bases ne sont pas constructivistes, de graves disjonctions pourraient facilement être négligées. Ainsi, par exemple, tirer des données d'une étude conçue en termes positivistes confère à ces données une valeur de vérité, un caractère absolu, qu'elles ne méritent pas en termes constructivistes. Dans un cadre constructiviste, ces mêmes données sont considérées comme variables et transformables, selon le point de vue du constructeur. L'utilisation de ces données positivistes dans le cadre d'une évaluation constructiviste sape l'essence même de l'évaluation. Les auteurs et les autrices de la littérature en évaluation, y compris les rapports d'évaluation, qui sont basés sur des principes constructivistes, feront presque certainement ressortir leur intention. Dans d'autres cas, l'apparition de concepts clés comme la généralisabilité, l'objectivité, la preuve, etc., typiques du positivisme et d'autres approches non constructives, peut être un signal clé quant à l'intention de l'auteur.

La phase d'assimilation de l'évaluation constructiviste représente l'effort de l'évaluateur pour intégrer les nouvelles découvertes dans la ou les constructions existantes (ou, si la nouvelle découverte est suffisamment différente de la ou des constructions existantes ou en conflit avec celles-ci, les remplacer) de sorte que la « nouvelle » construction (plus informée et plus sophistiquée) s'adapte (subsume les significations plus anciennes et plus récentes), fonctionne (explique ce qui se passe), démontre sa pertinence (permet de résoudre les problèmes fondamentaux, ou de les améliorer, ou de mieux les définir) et soit modifiable (ouverte elle-même au changement).

La découverte et l'assimilation ne sont pas nécessairement des processus séquentiels, mais peuvent se chevaucher ou être menées en parallèle.

Le processus d'évaluation constructiviste : les responsabilités de l'évaluateur constructiviste

L'évaluation constructiviste est un processus d'évaluation qui répond à deux conditions : elle est organisée en fonction des revendications, des préoccupations et des problèmes des publics concernés, et elle utilise la méthodologie du paradigme constructiviste. Compte tenu de ce constat, il est possible d'énumérer les neuf principales responsabilités dont l'évaluateur ou l'évaluatrice constructiviste doit s'acquitter. Il ou elle doit :

1. Identifier l'ensemble des parties prenantes qui sont à risque en raison des enjeux qu'elles détiennent dans l'entité faisant l'objet de l'évaluation. Ces enjeux peuvent comprendre, sans s'y limiter, l'argent, le statut, le pouvoir, les opportunités, ou autre; ces enjeux sont déterminés et définis par les parties prenantes (dans leurs propres termes) et pas seulement par l'évaluateur/-trice ou le/la client-e qui sollicite l'évaluation (bien qu'ils et elles soient eux aussi des parties prenantes et qu'ils et elles puissent identifier leurs propres enjeux et définitions). Les enjeux négatifs peuvent inclure l'exploitation, la déresponsabilisation et la privation des droits. Les parties prenantes ont le droit de recevoir et d'évaluer dans leurs propres termes toutes les informations que l'évaluation peut divulguer. Dans le processus herméneutique/dialectique qui s'ensuit, les différents enjeux qui entrent dans l'évaluation sont évalués et affinés dans le but de se rapprocher le plus possible de l'accord négocié. Il incombe à l'évaluateur de rechercher toutes les parties prenantes, y compris celles qui souhaitent rester discrètes ou s'absenter complètement.
2. Obtenir des groupes de parties prenantes leurs réflexions sur la forme et le processus de l'évaluanda et la gamme des revendications,

des préoccupations et des questions qu'ils souhaitent soulever à cet égard. La liste initiale peut être réorganisée, supprimée ou complétée au fur et à mesure de l'évaluation.

3. Fournir un contexte et une méthodologie (herméneutique/dialectique) par lesquels différentes constructions du sujet évalué et, différentes revendications, préoccupations et enjeux peuvent être compris, soumis à la critique et pris en compte. Le processus est d'abord mené au sein de groupes spécifiques de parties prenantes; puis les produits de ces négociations intragroupes (constructions définies, revendications, préoccupations et enjeux) sont négociés dans des cercles herméneutiques qui transcendent les groupes de parties prenantes, si nécessaire, dans des contextes dialogiques, contradictoires ou conflictuels.
4. Susciter un consensus sur le plus grand nombre possible de constructions, ainsi que sur les revendications, préoccupations et questions connexes. Le consensus devrait d'abord être recherché à l'intérieur d'un groupe, puis entre les groupes. S'il est possible de parvenir à un consensus sur un point, celui-ci peut être éliminé de la discussion, mais conservé pour une action ultérieure (et son inclusion dans le rapport d'évaluation) s'il y a accord sur celle-ci.
5. Préparer un ordre du jour pour la négociation des points sur lesquels il n'y a pas de consensus ou un consensus incomplet. L'absence de consensus implique la poursuite de constructions concurrentes, dont la ou les disjonctions ne peuvent être améliorées que par l'introduction de nouvelles informations ou par une augmentation du niveau de sophistication analytique. La tâche de l'évaluateur est d'identifier les informations nécessaires. Étant donné qu'il peut être nécessaire d'obtenir plus d'informations qu'il n'est possible d'en collecter, compte tenu des contraintes de temps et/ou de ressources, l'évaluateur doit trouver un moyen (de préférence également par le biais d'un processus herméneutique/dialectique) de hiérarchiser les éléments qui posent problème. Les contributions des parties prenantes sont essentielles dans cette détermination, de peur que ce besoin ne soit pris comme une opportunité de

déresponsabiliser certaines parties prenantes.

6. Collecter et fournir les informations demandées dans l'ordre du jour des négociations. La fourniture des informations nécessaires ne peut être garantie, mais l'évaluateur doit faire tout ce qui est en son pouvoir pour y parvenir. En outre, si les parties prenantes n'ont pas les connaissances nécessaires pour traiter les informations obtenues, une formation doit être dispensée et organisée par l'évaluateur.
7. Établir et arbitrer un forum de représentants des parties prenantes au sein duquel les négociations peuvent avoir lieu. Les différences non résolues dans les constructions, ainsi que les revendications, préoccupations et questions non résolues sont examinées à la lumière des nouvelles informations et du niveau de sophistication, dans l'espoir que leur nombre puisse être réduit. Il est probable que certains points ne seront pas résolus, ce qui ouvrira la voie à une autre série d'activités d'évaluation ultérieures. Les résultats de ce forum doivent comprendre des actions à prendre si l'on veut que la négociation soit considérée comme un succès.
8. Élaborer un rapport, probablement plusieurs rapports ciblés, qui communiquent à chaque groupe de parties prenantes tout consensus sur les constructions et toute résolution concernant les revendications, les préoccupations et les questions qu'ils et elles ont soulevées (ainsi que celles soulevées par d'autres groupes qui semblent pertinentes pour ce groupe). La forme la plus utile pour ce(s) rapport(s) est l'étude de cas, qui peut fournir l'expérience indirecte nécessaire pour influencer les constructions des parties prenantes (voir ci-dessous des observations supplémentaires sur le processus d'établissement de rapports).
9. Recycler l'évaluation pour tenir compte des constructions encore non résolues et de leurs revendications, préoccupations et questions connexes. De nouveaux aspects peuvent être explorés sur la base de l'évaluation initiale. Les évaluations constructivistes ne sont jamais achevées; elles s'arrêtent jusqu'à ce qu'un nouveau besoin ou une nouvelle occasion de révision et de réévaluation apparaisse.

[…]

Conduire l'évaluation constructiviste; l'utilisation de la méthodologie herméneutique/dialectique

L'évaluation constructiviste s'effectue par le biais d'une série d'étapes qui, tout en étant énumérées ici en série, peuvent être itératives et réitérées dans la pratique à mesure que les constructions évoluent et que des revendications, des préoccupations et des questions particulières sont traitées. La présentation en série ci-dessous est utilisée pour des raisons de commodité. La liste commence au moment où un contrat satisfaisant pour toutes les parties a été conclu.

1. Organiser l'évaluation : sélectionner l'équipe initiale d'évaluateurs et d'évaluatrices, prendre des dispositions d'entrée, prendre des dispositions logistiques et évaluer les facteurs politiques et culturels locaux.
2. Identifier les parties prenantes : identifier les agents et les agentes qui commanditent et réalisent l'évaluanda, identifier les « bénéficiaires » ainsi que les « victimes » de l'action du sujet évalué, monter des stratégies de recherche continue pour d'autres parties prenantes, évaluer les compromis et les sanctions, et formaliser les accords avec et entre elles et eux.
3. Développer des constructions intragroupes de parties prenantes : former plusieurs cercles herméneutiques de 10 à 12 membres représentant chacun un public de parties prenantes; solliciter des descriptions (constructions) de l'évaluanda et identifier et approfondir les revendications, les préoccupations et les questions qui émergent, pour aboutir autant que faire se peut à des accords négociés sur tous les points identifiés.
4. Élargir les constructions conjointement élaborées par les groupes de parties prenantes en utilisant les constructions antérieures de

l'évaluateur ou de l'évaluatrice (mais sans leur accorder de privilège particulier), les informations documentaires existantes, une mise en regard des données issues des entretiens intragroupes avec les données d'observation, l'analyse de la littérature et d'autres sources jugées pertinentes.

5. Trier les constructions, les revendications, les préoccupations et les questions résolues par consensus, en les mettant de côté comme éléments possibles dans les rapports de cas.
6. Hiérarchiser les points non résolus au moyen d'un processus de hiérarchisation négocié, déterminé par les membres de chaque groupe de parties prenantes et impliquant des derniers.
7. Recueillir des informations supplémentaires et s'en servir pour ajouter de la sophistication en formant les négociateurs et les négociatrices, en recherchant de nouvelles informations, en réalisant des études spéciales si nécessaire.
8. Préparer l'ordre du jour des négociations en définissant et en élucidant les constructions concurrentes; en s'efforçant d'éclairer, de soutenir ou de réfuter les éléments (en fournissant une formation supplémentaire si nécessaire) et en testant l'ordre du jour obtenu.
9. Développer des constructions inter-groupes. L'étape 8 aura abouti à un ordre du jour négocié pour chacun des groupes de parties prenantes. Cette étape 9 récapitule en fait les étapes 3 à 8 pour un cercle herméneutique nouvellement formé, composé de personnes choisies par les différents cercles comme leurs représentantes et représentants. Le résultat est une construction composite qui inclut toutes les formes de constructions du sujet évalué ainsi que leurs revendications, préoccupations et questions pertinentes. Il est pratiquement certain que certains points n'auront pas été négociés jusqu'à ce que tous les groupes de parties prenantes soient satisfait-e-s; ceux-ci/celles-ci sont mis-e-s de côté en vue d'un réexamen ultérieur dans le cadre d'un recyclage.
10. Rapport sur les résultats de l'étape 9. Il peut y avoir plusieurs rapports adaptés aux revendications, aux préoccupations et aux questions de groupes de parties prenantes spécifiques. Les accords sur les

éléments de ces rapports peuvent mener à des propositions de mesures à prendre. Le rapport devrait viser en particulier le(s) objectif(s) de l'évaluation, c'est-à-dire formatif/mérite, formatif/valeur, sommatif/mérite, et/ou la valeur sommative.

11. Recycler l'ensemble du processus pour tenir compte en particulier des éléments mis de côté à l'étape 9 et qui n'étaient pas résolubles à l'époque.

Rapports d'évaluation constructiviste

Le rendu d'une évaluation constructiviste (mais jamais son produit final, puisqu'il est soumis à des itérations successives) est le rapport de cas. En un sens, une étude de cas n'est jamais terminée, elle est simplement due. Il peut y avoir de multiples rapports destinés à des publics de parties prenantes spécifiques; et ils peuvent prendre de nombreuses formes, possiblement sans y inclure ce que l'on pourrait normalement appeler un rapport « technique », si un tel rapport dépasse les compétences d'un public de parties prenantes. Le rapport n'aboutit pas à des jugements, des conclusions ou des recommandations, sauf dans la mesure où ceux-ci sont approuvés par les parties prenantes concernées.

Le rapport de cas est plutôt la construction conjointe qui émerge du processus herméneutique/dialectique. Tout au long de ce processus, les parties prenantes – individuellement, dans des groupes homogènes ou des groupes dans lesquels elles sont mélangées – sont choisies pour découvrir des points de vue très différents. Elles sont exposées à de nouvelles informations et à de nouvelles méthodes d'analyse et d'interprétation plus sophistiquées jusqu'à ce qu'un certain consensus soit atteint.

Le rapport de cas aide le lecteur à comprendre (dans le sens de rendre réel), non seulement l'état des choses que les parties prenantes croient exister, mais aussi les motifs sous-jacents, les sentiments et les raisons

qui mènent à ces croyances. Le rapport de cas se caractérise par une description détaillée qui non seulement clarifie le contexte essentiel, mais qui permet au lecteur d'en faire l'expérience par procuration.

Le rapport de cas doit enfin contenir une annexe qui décrit en détail la méthodologie suivie et permet de juger dans quelle mesure les critères de qualité (ceux énumérés dans la section suivante) sont remplis.

Critères d'évaluation de la qualité des évaluations et des rapports constructivistes

Les normes habituellement appliquées pour juger de la qualité des évaluations, par exemple, les normes du *Joint Committee* ou *les Guiding Principles for Evaluators* [Principes directeurs pour les évaluateurs] de l'*American Evaluation Association*, ne conviennent pas aux évaluations constructivistes, précisément parce qu'elles reposent sur un paradigme théorique fondamentalement différent (comme expliqué dans les premiers paragraphes de ce texte). Deux approches différentes ont été élaborées pour faire face à ce dilemme; toutes deux sont utiles pendant le processus d'évaluation en tant que listes de contrôle procédurales et, par la suite, pour évaluer la qualité du rapport d'évaluation (produit) :

1. Les critères « parallèles » (parfois appelés « critères de fiabilité » ou « fondamentaux »). Ils sont le fruit d'un effort visant à produire des critères plus ou moins parallèles à ceux utilisés de manière conventionnelle, c'est-à-dire la validité interne et externe, la fiabilité et l'objectivité. Ils sont probablement les plus utiles, d'abord pour orienter les décisions méthodologiques pendant l'évaluation, puis pour vérifier l'ensemble du processus d'évaluation (voir c et d ci-dessous). Cependant, leur « parallélisme » avec les principes positivistes ne les rend pas tout à fait adéquats pour déterminer la qualité d'une approche constructiviste. Ces critères parallèles sont

les suivants (on trouvera des définitions complètes dans *Fourth Generation Evaluation*, p.233 à 243) :

2. Crédibilité, à peu près parallèle à la validité interne, établie par un engagement prolongé sur le site, une observation persistante, un compte rendu par les pairs (une sorte de critique externe), une analyse des cas négatifs (un processus de reformulation des hypothèses), une subjectivité progressive (vérification continue des constructions en cours de développement par rapport aux constructions attendues avant la collecte des données) et (plus important encore), des vérifications par les membres, une évaluation continue des hypothèses, des données, des catégories préliminaires, et des interprétations avec les membres des publics concernés.
3. Transférabilité, à peu près parallèle à la validité externe, établie non pas par l'évaluateur ou l'évaluatrice, mais par les destinataires des rapports d'évaluation qui jugent personnellement la mesure dans laquelle les résultats sont suffisamment similaires à leurs propres situations (à partir de la description détaillée) pour justifier la vérification de la viabilité de l'application locale (vérification de la localisation plutôt que la généralisation plus habituelle).
4. Solidité [*Dependability*], à peu près parallèle à la fiabilité [*reliability*], établie grâce à l'audit de solidité avec l'aide d'une auditrice ou d'un auditeur externe, qui examine le dossier de l'enquête de la façon dont un auditeur fiscal examine les dossiers fiscaux, afin de déterminer les décisions méthodologiques prises et d'en comprendre les raisons.
5. Confirmabilité, à peu près parallèle à l'objectivité, qui détermine dans quelle mesure les constructions, les affirmations, les faits et les données peuvent être retracés jusqu'à leurs sources, l'inspection étant effectuée par un auditeur ou une auditrice externe (qui peut être identique ou différent-e de l'auditeur/-trice de la solidité). Les « produits bruts » et les « processus utilisés pour les comprimer » sont inspectés et confirmés selon qu'il convient.
6. Les critères d'authenticité. Alors que les critères parallèles sont ancrés dans les hypothèses de positivisme, les critères d'authenticité sont fondés directement sur les hypothèses de constructivisme et

répondent aux aspects herméneutiques/dialectiques de ce paradigme. Ces critères sont les suivants (les définitions complètes se trouvent dans *Fourth Generation Evaluation*, p.245 à 250) :

7. Équité, déterminée par une évaluation de la mesure dans laquelle toutes les constructions concurrentes ont été accessibles, exposées et prises en compte dans le rapport d'évaluation, c'est-à-dire dans la construction émergente négociée.
8. Authenticité ontologique, déterminée par une évaluation de la mesure dans laquelle les constructions individuelles (y compris celles de l'évaluateur ou de l'évaluatrice) sont devenues plus informées et plus sophistiquées.
9. Authenticité éducative, déterminée par une évaluation de la mesure dans laquelle les individus (y compris l'évaluateur ou l'évaluatrice) comprennent mieux (même s'ils ou elles ne sont pas plus tolérant-e-s) les constructions des autres.
10. Authenticité catalytique, déterminée par une évaluation de la mesure dans laquelle l'action est stimulée et facilitée par l'évaluation (en clarifiant le centre d'intérêt, en éliminant ou en améliorant les problèmes, en affinant les valeurs).
11. Authenticité tactique, déterminée par une évaluation de la mesure dans laquelle les individus sont habilités à prendre les mesures que l'évaluation implique ou propose.

Deux autres observations s'imposent en ce qui concerne la question de la qualité. Premièrement, il ne faut pas négliger la capacité du processus herméneutique/dialectique à agir comme une source puissante de contrôle de la qualité. Dans ce processus, les données saisies sont analysées dès leur réception. Elles sont « renvoyées », pour être commentées, élaborées, corrigées, révisées, développées ou modifiées, aux personnes qui les ont fournies quelques instants avant. Ces intrants seront en outre intégrés à la reconstruction conjointe et collaborative qui émerge à mesure que le processus se poursuit. Dans ces circonstances, les possibilités que des erreurs ne soient pas détectées ou contestées sont très faibles. C'est l'interaction immédiate et continue des informations

qui empêche la possibilité de résultats non crédibles. Il est difficile de maintenir une façade factice ou de soutenir une tromperie délibérée lorsque l'information est soumise à des remises en question continues et multiples de la part de diverses parties prenantes. Le processus herméneutique/dialectique est par définition susceptible d'être inspecté et est en effet inspecté lui-même; ceci empêche une grande partie des types de secrets et de pauvreté de l'information qui ont caractérisé les évaluations se concentrant uniquement sur les attentes des clients [*client-focused evaluations*]. Enfin, toute intention de la part de l'évaluateur ou de l'évaluatrice de favoriser certaines parties prenantes est au moins tout aussi détectable.

Deuxièmement, pour qu'une évaluation de la qualité aboutisse, il est nécessaire que l'évaluatrice ou l'évaluateur joue un double rôle (et parfois contradictoire) : défenseur ou défenseuse et éducateur ou éducatrice. Dans pratiquement toutes les situations, les publics intéressés diffèrent grandement quant à la quantité d'informations qu'ils et elles fournissent, la mesure dans laquelle ils et elles peuvent articuler leurs constructions existantes du sujet évalué et les revendications, préoccupations et problèmes auxquels ils et elles sont confronté-e-s; et le degré de sophistication qu'ils et elles possèdent dans le traitement des nouvelles informations qui émergent, dont certaines peuvent être très techniques. En outre, le processus herméneutique/dialectique lui-même n'est pas un processus qu'ils et elles maîtrisent bien; il incombe donc à l'évaluateur ou à l'évaluatrice de leur fournir la formation (et, si nécessaire, la représentation) dont ils et elles ont besoin. L'équilibre requis entre ces rôles est délicat, et l'évaluateur ou l'évaluatrice devra faire preuve d'une grande prudence pour éviter tout parti pris et tout favoritisme.

Coda

L'évaluation constructiviste diffère fondamentalement des autres (nombreuses) formes d'évaluation. Dans *Fourth Generation Evaluation*, nous avons décrit l'évolution historique de la pratique de l'évaluation : une première génération était axée sur la mesure, une deuxième sur la description, une troisième sur le jugement et une quatrième sur la négociation (herméneutique/dialectique). C'est cette quatrième forme d'évaluation qui fait l'objet de cette liste de contrôle et de cet ensemble de lignes directrices, maintenant appelée évaluation constructiviste. Nous pensons que cette forme élimine les problèmes majeurs des trois premières générations : une tendance au managérialisme, c'est-à-dire une approche de l'évaluation qui favorise le point de vue du client ou de la cliente, du bailleur ou de la bailleuse de fonds, qui préserve indûment le ou la gestionnaire et qui prive certaines parties prenantes de leur pouvoir, de leur équité et de leurs droits; une incapacité à tenir compte du pluralisme des valeurs; et un engagement excessif envers le paradigme scientifique (positiviste) de l'enquête.

L'évaluation constructiviste est un modèle difficile à adopter. Elle demande beaucoup de travail. Elle est toujours récursive et nécessite de fréquentes récapitulations. Elle est souvent contradictoire et conflictuelle. Il s'agit d'un processus diffus impossible à spécifier en détail (sous forme de modèle); par conséquent, ses engagements en matière de personnel et de ressources peuvent au mieux être « estimés ». Ceci exige de l'évaluateur ou de l'évaluatrice qu'il ou elle joue de multiples rôles qui peuvent parfois sembler contradictoires. Elle nie la possibilité de généralisations fiables et de trouver des solutions « qui fonctionnent » partout. Pourtant, d'un point de vue axé sur les valeurs, c'est, selon nous, le meilleur moyen d'élaborer des solutions viables et acceptables aux revendications, aux préoccupations et aux problèmes largement ressentis, et à la formulation de constructions largement considérées

comme adaptées, efficaces, pertinentes et modifiables en permanence. C'est l'une des approches les plus réalistes sur le plan social et politique pour réaliser des évaluations utiles et utilisées.

[…]

5. Une approche fondée sur les valeurs pour évaluer le projet scolaire Bunche-Da Vinci

JENNIFER C. GREENE

[Traduit de : Greene, Jennifer C. 2005. « A value-engaged approach for evaluating the Bunche-Da Vinci Learning Academy ». *New Directions for Evaluation*, 2005(106) : 27-45 (Extraits). Traduction par Carine Gazier et Thomas Delahais; traduction et reproduction du texte avec l'autorisation de John Wiley and Sons.]

Une approche de l'évaluation de programme fondée sur les valeurs combine des éléments issus de l'évaluation répondante ou *responsive evaluation* (Abma et Stake, 2001; Stake 1987, 2004) avec un engagement actif envers des valeurs qui sont tirées principalement de traditions évaluatives attentives aux enjeux démocratiques et aux différences culturelles (Hood, 1999; House et Howe, 1999; MacDonald, 1977). Ces valeurs comprennent un engagement à prendre en compte le contexte, l'inclusion des parties prenantes, l'apprentissage, la diversité et le bien public (comme nous le verrons plus loin dans ce chapitre). Étant donné que l'évaluation fondée sur les valeurs est fondamentalement sensible au caractère particulier d'un contexte d'évaluation et aux questions programmatiques et politiques présentes dans ce contexte, il n'est pas possible d'élaborer un modèle d'évaluation a priori (Stake, 1987) en utilisant cette approche. Au contraire, le modèle de l'évaluation évolue au fur et à mesure que l'évaluateur ou l'évaluatrice apprend à connaître le contexte, ses acteurs et actrices, ses joies et ses difficultés. Dans cette approche de l'évaluation, des efforts considérables sont consacrés aux aspects initiaux de la pratique de l'évaluation : apprendre à connaître le

contexte et le programme à évaluer, établir des relations appropriées avec les principales parties prenantes, comprendre en profondeur les points critiques qui devront être abordés, identifier les questions d'évaluation prioritaires et déterminer les critères permettant de juger de la qualité du programme. Ce n'est qu'une fois que tous ces aspects de la pratique d'évaluation auront été pris en compte que l'on pourra élaborer et mettre en œuvre un modèle d'évaluation. Cette approche rend le processus évaluatif intimement lié au contexte.

Par conséquent, ce récit de l'évaluation du projet scolaire Bunche-Da Vinci à partir d'une approche fondée sur les valeurs se concentre sur la façon dont nous avons considéré ces aspects initiaux de l'évaluation dans ce contexte particulier. Des hypothèses seront également formulées en cours de route pour permettre de raconter une histoire plus complète. […]

La communauté éducative Bunche-Da Vinci

La communauté de l'école élémentaire Bunche-Da Vinci ressemble à de nombreuses communautés scolaires urbaines d'aujourd'hui, tant au regard des données démographiques qui montrent une surreprésentation des minorités raciales et ethniques aux États-Unis, d'un historique d'insuffisance des ressources affectées à l'éducation, et de défis persistants en matière d'efficacité de l'enseignement et de l'apprentissage. Plus précisément, la communauté scolaire Bunche-Da Vinci est située dans une zone géographique tampon entre le secteur industriel d'une ville portuaire et une banlieue historiquement ouvrière, connue pour sa pauvreté, sa criminalité et ses tensions raciales. L'école attire principalement des élèves issus de familles économiquement marginales de ces communautés voisines – à l'heure actuelle, les Latino-Américains et Latino-Américaines sont majoritaires (près de 80 %) et s'y ajoutent 17 % d'Afro-Américains et Afro-Américaines, sur une population scolaire

totale d'environ 1 200 élèves. Les effectifs latino-américains ont augmenté de 50 % au cours des dix dernières années, avec une diminution concomitante dans tous les autres groupes. L'anglais est la deuxième langue pour près des deux tiers des élèves; l'espagnol étant pour presque tou-te-s leur première langue. Le renouvellement des élèves d'une année sur l'autre à Bunche-Da Vinci est considérable, même si de nombreux parents expriment leur satisfaction à l'égard de l'école et que la plupart des élèves semblent aimer être dans cette école.

Une première étape très importante dans l'élaboration d'un plan d'évaluation pour l'école Bunche-Da Vinci consiste à mieux comprendre les caractéristiques et les diversités particulières de cette communauté scolaire. Étant donné que cette communauté scolaire partage de nombreuses caractéristiques avec d'autres communautés scolaires urbaines d'aujourd'hui, il est particulièrement important de *ne pas* faire d'hypothèses sur ce contexte particulier à partir de stéréotypes urbains. On pourrait ainsi supposer, par exemple, que les parents Bunche-Da Vinci sont au mieux modestement impliqués dans l'éducation de leurs enfants ou que la drogue domine les échanges économiques de la communauté. Apprendre à connaître cette communauté scolaire spécifique – son caractère unique, ses complexités et son évolution continue et dynamique – est fondamental pour développer un modèle d'évaluation susceptible de générer des informations significatives. À cette fin, il se peut que l'évaluatrice ou l'évaluateur doive recruter, en tant que membres de l'équipe d'évaluation ou en tant que consultants ou consultantes, des personnes qui ont un passé socioculturel ou politique partagé avec les membres de la communauté scolaire Bunche-Da Vinci. Les informateurs et les informatrices clés au sein de l'école peuvent aider à en apprendre davantage sur cette communauté, mais cela ne sera probablement pas suffisant. […]

Le programme de réforme scolaire Da Vinci

Il y a trois ans, l'école Bunche a été choisie par son district pour s'associer à l'organisation Da Vinci pour lancer une initiative ambitieuse de réforme globale de l'école, spécialement conçue par cette organisation pour les écoles peu performantes. Da Vinci a le contrôle sur les aspects essentiels des programmes de l'école et le contrôle total de son budget, mais l'école continue de faire partie du district scolaire. [Ce que cet arrangement] signifie en réalité n'est pas tout à fait clair.

Au regard des informations disponibles, le partenariat semble avoir connu des tensions et des défis depuis sa création il y a trois ans. [...] En particulier, les enseignants et les enseignantes, dont bon nombre sont assez jeunes et environ un quart n'ont pas suivi de formation à l'enseignement, signalent être en situation de stress et être mis sous une pression considérable pour qu'ils et elles travaillent de longues heures à la mise en œuvre d'un programme qui leur semble étranger ou « importé », et dont la plupart ne sont pas « comptées » en vertu de la loi fédérale *No Child Left Behind*[1] [Aucun enfant laissé pour compte]. Par exemple, le programme d'apprentissage de la lecture Da Vinci diffère de celui adopté par l'État, et les exigences technologiques du programme semblent prendre un temps précieux sur l'enseignement et l'apprentissage de base. Certain-e-s enseignant-e-s ont même laissé entendre qu'ils et elles ont l'impression que leur jugement professionnel est sapé par la structure rigide et les exigences du programme Da Vinci. La rotation des enseignant-e-s de l'école est élevée. La participation des parents à la gestion de l'école et aux activités scolaires est minime. Et les résultats des élèves aux tests de langues obligatoires de l'État, bien qu'ils aient

1. NdT : Cette loi américaine de 2001 institue l'obligation pour les écoles états-uniennes d'atteindre des standards de performance fixés et mesurés par les États fédérés, tant en termes d'organisation (compétences des enseignants, nombre d'heures assurées...) que de résultats des élèves. Seules les écoles faisant passer annuellement des tests standardisés aux élèves ont accès aux financements fédéraux.

légèrement progressé au cours de la deuxième année du partenariat, ont chuté à des niveaux inférieurs aux résultats de l'école au début du partenariat, à la fois globalement et pour tous les sous-groupes. Seul un quart environ de la population de l'école fait preuve de compétences dans ce test de langue (Les performances des étudiant-e-s aux tests de l'État dans d'autres matières pourraient bien être similaires). Et ce alors même que les performances des élèves aux tests standardisés Da Vinci se sont améliorées de manière constante au cours des trois dernières années. En outre, les parents et les élèves se disent satisfaits de l'école, ce qui n'est pas compatible avec le fait qu'un grand nombre d'élèves quittent l'école chaque année.

Qu'est-ce qui est évalué au juste?

Tout comme pour le développement d'une compréhension de la communauté scolaire au sens large, une compréhension plus complète du programme Da Vinci et de sa mise en œuvre à l'école Bunche est une première étape essentielle dans une approche de l'évaluation fondée sur les valeurs. Ce n'est qu'avec une telle compréhension que l'évaluation pourra être ancrée dans les préoccupations et les enjeux (Guba et Lincoln, 1989; Stake, 2004) qui sont prioritaires dans ce contexte. L'identification du périmètre d'évaluation, voire de ce qui est évalué tout simplement, n'est pas présumée dans cette approche d'évaluation, mais est plutôt un sujet de discussion entre autant de groupes de parties prenantes que possible dans cette partie initiale de l'évaluation. L'évaluateur ou l'évaluatrice apporte à cette discussion son expertise et ses perspectives, y compris les idées tirées de la littérature académique pertinente. Dans ce contexte, par exemple, les principes d'une évaluation basée sur la théorie peuvent être pertinents et utiles, car ils pourraient catalyser une explication des logiques conceptuelles sous-jacentes au programme Da Vinci et au partenariat Bunche-Da Vinci, et ainsi permettre l'examen et la critique de ces logiques dans le cadre de l'évaluation.

Ce contexte particulier se distingue notamment par l'intervention mise en place par la Société Da Vinci pour les écoles peu performantes, mais aussi par un partenariat public-privé tout à fait spécial. Ce sont tous deux des candidats viables pour une évaluation. Et tous deux abordent potentiellement des questions d'éducation de tout temps très importantes – questions de pédagogie, de programme, de différenciation et d'individualisation – ainsi que des questions plus spécifiques à notre époque : comment et quoi enseigner à des élèves en cours d'apprentissage de la langue anglaise, mais aussi des interrogations sur les contributions de la technologie à l'apprentissage, le professionnalisme des enseignant-e-s, ainsi que l'efficacité éducative des normes imposées par l'État, les exigences de redevabilité et la privatisation de l'enseignement public. Deux questions sont abordées dans ce processus d'évaluation préliminaire : quels enjeux importent dans ce contexte? Quelle forme particulière prennent-ils?

Consacrer un temps d'évaluation précieux à l'élaboration d'une compréhension précise et réfléchie de ce qui est évalué dans ce contexte particulier appuie l'élaboration de questions d'évaluation significatives et potentiellement utiles. Cela reflète également certains engagements en termes de valeurs, notamment :

- Un engagement à prendre en compte le contexte, pour comprendre la nature du programme (et/ou de la structure institutionnelle, comme un partenariat) à évaluer dans son contexte particulier et unique.
- Un engagement à inclure dans l'évaluation tous les points de vue et perspectives légitimes des parties prenantes, en s'efforçant tout particulièrement d'inclure les personnes les plus marginalisées dans le contexte de l'évaluation (House et Howe, 1999) – par exemple, dans cette communauté scolaire particulière, les enseignant-e-s découragé-e-s et les familles sans domicile fixe.

Dans le modèle d'évaluation fondée sur les valeurs qui sera élaboré dans ce contexte, un espace considérable est alloué à la description de ce qui est évalué et de ses caractéristiques distinctives. Cette description fournit l'ancrage contextuel et la justification du modèle d'évaluation qui est élaboré. [...]

L'objectif principal de l'évaluation dans le cadre d'une approche fondée sur les valeurs est d'accroître notre compréhension de la qualité et de l'efficacité du sujet évalué dans le contexte particulier qui nous occupe. Il s'agit d'une vision fondamentalement éducative de l'évaluation (Cronbach *et al.*, 1980). En outre, cela exige que l'évaluation comprenne des analyses du programme tel qu'il a été mis en œuvre et vécu, ainsi que des analyses des changements significatifs que le programme a apportés dans la vie de ses participants. De plus, dans une approche fondée sur les valeurs, les questions d'évaluation portent spécifiquement sur les intérêts des personnes traditionnellement mal considérées dans notre pays, y compris les minorités raciales et ethniques, les personnes à faible revenu, les immigrant-e-s et les personnes en situation de handicap. Les intérêts de la majorité ne sont pas exclus dans cette approche d'évaluation, mais il s'agit de porter une attention spécifique aux intérêts des minorités. Ce cadrage des questions d'évaluation reflète d'autres engagements de valeurs qui sous-tendent cette approche :

Un engagement à apprendre dans le cadre de l'évaluation et par l'entremise de celle-ci, afin de contribuer à une meilleure compréhension de la façon dont le programme évalué répond à des besoins et à des intérêts importants des participant-e-s dans ce contexte particulier (en plus de la performance des participant-e-s au programme).

Un engagement à s'intéresser à la différence et la diversité telles qu'elles se manifestent dans le contexte en question, afin d'évaluer dans quelle mesure le programme évalué répond aux besoins et aux intérêts

particuliers des personnes traditionnellement mal considérées, dont pourraient bien faire partie la plupart des enfants et des familles dans le contexte de Bunche-Da Vinci.

Compte tenu de cette démarche en vue d'identifier les questions d'évaluation, et des priorités d'évaluation identifiées par les parties prenantes de Bunche-Da Vinci, les principales questions d'évaluation à aborder [...] sont (provisoirement) les suivantes :

Question générale :

1. De quelle manière et dans quelle mesure le programme éducatif proposé à l'Académie Bunche-Da Vinci répond aux besoins éducatifs importants des enfants et des familles dépendant-e-s de cette école, en particulier les besoins particuliers des élèves en cours d'apprentissage de l'anglais, des enfants issus de groupes raciaux et ethniques minoritaires, des enfants issus de familles à faible revenu et des enfants ayant des besoins spéciaux? Et de quelle manière et dans quelle mesure la structure du partenariat soutient-elle cette mission éducative principale? [...]
2. Dans quelle mesure et de quelle manière les enfants de l'Académie Bunche-Da Vinci *obtiennent-ils et elles des résultats scolaires significatifs et valorisés*? [...]

Établir les critères permettant de juger de la qualité du programme

Une dernière facette initiale de l'évaluation concerne les critères à utiliser pour juger de la qualité du programme. Dans le cadre d'une approche fondée sur les valeurs, ces critères ne sont pas supposés, mais plutôt établis par des discussions avec diverses parties prenantes, en tandem avec des perspectives externes pertinentes fournies par l'évaluateur ou l'évaluatrice. En outre, les discussions sur les critères de qualité sont

particulièrement importantes pour l'inclusion des parties prenantes, car les critères peuvent et doivent impliquer et refléter la multiplicité des valeurs et des idéaux qui leur sont chers dans le contexte en question. [...]

- Les critères pour juger de la qualité de l'enseignement dans ce contexte sont compliqués par la présence du programme Da Vinci, qui a probablement ses propres critères de qualité (qui ne sont pas encore connus), et par la diversité des parties prenantes du programme. Une ébauche des critères généraux envisagés comme possibles dans ce contexte comprend les éléments suivants :
- Le programme d'études et la pédagogie sont de grande qualité, fondés sur une théorie et des recherches pertinentes (par exemple, des recherches portant sur les enseignant-e-s efficaces avec des enfants issus de minorités) et sur les perspectives et les expériences du personnel et des familles de Bunche-Da Vinci.
- Le programme d'études et la pédagogie offrent des approches d'apprentissage pertinentes et valables pour l'ensemble de la population des étudiants de Bunche-Da Vinci.
- Le programme d'études et ses évaluations sont bien alignés sur les normes de l'État dans toutes les matières, en particulier les langues et les mathématiques.
- Tou-te-s les élèves de Bunche-Da Vinci ont des expériences d'apprentissage significatives et positives et font preuve d'une maîtrise solide et constante des compétences et des connaissances valorisées, en particulier en langues et en mathématiques, telles qu'évaluées par les tests Da Vinci, les tests d'État et le jugement des enseignant-e-s, parmi d'autres mesures possibles des résultats. Aux tests de l'État, les performances des élèves doivent être suffisamment élevées chaque année pour que l'école ne soit pas inscrite sur la « liste de surveillance » de l'État.
- La communauté éducative de l'école Bunche-Da Vinci nourrit de grandes espérances pour l'apprentissage de tou-te-s les élèves, valorise et affirme la valeur de tous les membres de la communauté, et fait preuve de bienveillance et de soutien envers tous les membres

de communauté; *et* les membres de cette communauté sont heureux et satisfaits d'en faire partie.
- La structure du partenariat Bunche-Da Vinci affirme et soutient la responsabilité et l'autorité publiques en matière d'éducation des enfants de cette communauté.
- Le dernier critère de cette liste représente un engagement de valeur final dans l'évaluation fondée sur les valeurs, plus précisément :
- Un engagement à aborder les questions liées au bien public dans un contexte d'évaluation, c'est-à-dire les questions liées à la responsabilité civique, aux priorités en matière de politiques publiques, aux idéaux démocratiques ou, dans ce contexte, aux responsabilités publiques vitales en matière d'éducation. [...]

Le processus d'évaluation

Une approche de l'évaluation fondée sur les valeurs se fait au plus près, à l'intérieur et autour de l'environnement du sujet évalué. Il n'y a pas d'autre moyen de développer une compréhension précise et sensible des caractéristiques uniques du sujet évalué dans son contexte particulier. De plus, une approche fondée sur les valeurs répond fondamentalement aux questions et aux préoccupations de tous les groupes de parties prenantes légitimes dans ce contexte. Cela nécessite également une présence sur le terrain et une écoute attentive des rythmes multiples et variés des expériences vécues dans cette communauté particulière. Une approche fondée sur les valeurs est également fondamentalement éducative, car elle vise à fournir des espaces et des lieux pour des réflexions approfondies et fondées sur des données de la pratique – dans ce cas, sur les programmes scolaires, l'enseignement, l'apprentissage et les besoins éducatifs particuliers de divers types d'enfants.

Compte tenu de ces caractéristiques, l'évaluateur ou l'évaluatrice engagé-e doit s'intéresser non seulement aux dimensions matérielles et méthodologiques de l'évaluation, mais aussi à ses dimensions relationnelles, à la façon dont l'évaluateur/-trice est présente dans le contexte en question – c'est-à-dire au processus d'évaluation, à l'évaluation elle-même en tant que pratique morale et éthique (Schwandt, 2004). Dans le contexte de Bunche-Da Vinci, ces dimensions relationnelles et pratiques de l'évaluation seront délibérément mises en œuvre de trois manières principales. Premièrement, l'évaluateur/-trice s'efforcera d'établir des relations de respect, de réciprocité et de partenariat, avec toutes les parties prenantes de la communauté scolaire Bunche-Da Vinci. Le respect et la réciprocité peuvent être communiqués, par exemple, en supposant que toutes les parties prenantes ont quelque chose de précieux à apporter à l'évaluation et en écoutant attentivement chacune d'elles (Greene, 2003). Deuxièmement, le processus d'évaluation sera ouvert et transparent, et invitera de multiples parties prenantes à participer à toutes les phases, mais surtout au début, lorsque le programme d'évaluation est établi, et à la fin, lorsque les interprétations des résultats et les implications des actions sont discutées. Le plus important en ce qui a trait à la participation est l'inclusion de toutes les perspectives légitimes des parties prenantes dans l'évaluation et l'ancrage du travail d'évaluation dans les particularités de ce contexte. La participation effective des parties prenantes aux activités concrètes de collecte et d'analyse des données est moins importante, bien que les parties prenantes soient également les bienvenues dans ces processus. Et troisièmement, l'évaluateur/-trice engagé-e cherchera des occasions de dialogue constructif avec les parties prenantes, en particulier des occasions pour celles-ci de dialoguer entre elles sur leurs expériences et perceptions respectives du projet éducatif Bunche-Da Vinci (Abma, 2001). Bien que de telles possibilités puissent être rares, l'évaluateur/-trice se fondant sur les valeurs considère le dialogue comme un moyen particulièrement précieux d'instaurer une relation respectueuse, la réactivité, et surtout l'apprentissage. […]

Bibliographie

Abma, Tineke A. 2001. « Special Issue: Dialogue in Evaluation ». *Evaluation* 7.

Abma, Tineke A., et Robert E. Stake. 2001. « Stake's Responsive Evaluation: Core Ideas and Evolution ». in *Responsive Evaluation. New Directions for Evaluation*, N°92, édité par J. C. Greene et T. A. Abma. San Francisco: Jossey-Bass.

Cronbach, Lee J., Sueann R. Ambron, Sanford M. Dornbusch, Robert D. Hess, Robert C. Hornik, Denis Charles Phillips, Decker F. Walker et Stephen S. Weiner. 1980. *Toward Reform of Program Evaluation: Aims, Methods, and Institutional Arrangements*. San Francisco: Jossey-Bass.

Greene, Jennifer C. 2003. « Evaluators as Stewards of the Public Good ». Présenté à Quatrième conférence sur la pertinence et la culture de l'évaluation, Tempe, Ariz.

Guba, Egon G., et Yvonne S. Lincoln. 1989. *Fourth generation evaluation*. Thousand Oaks: Sage Publications.

Hood, Stafford. 1999. « Responsive Evaluation Amistad Style: Perspectives of One African-American Evaluator ». in *Proceedings of the 1998 Robert E. Stake Symposium on Educational Evaluation*, édité par R. Davis. Urbana-Champaign: University of Illinois.

House, Ernest R., et Kenneth R. Howe. 1999. *Values in Evaluation and Social Research*. 1re éd. Thousand Oaks: Sage Publications.

MacDonald, Barry. 1977. « A Political Classification of Evaluation Studies ». in *Beyond the Numbers Game*. New York: Macmillan.

Schwandt, Peter. 2004. *Evaluation practice reconsidered*. New York: Peter Lang.

Stake, Robert. 1987. « Program Evaluation, Particularly Responsive Evaluation ». in *Program Evaluation, Particularly Responsive Evaluation*, édité par G. F. Madaus, M. Scriven et D. L. Stufflebeam. Norwood: Kluwer-Nijhoff.

Stake, Robert E. 2004. *Standards-Based and Responsive Evaluation*. Thousand Oaks: Sage Publications.

6. Accroître la compétence culturelle, une étape nécessaire en appui à une évaluation menée par les personnes autochtones

NAN WEHIPEIHANA

[Traduction de Wehipeihana, Nan. 2019. « Increasing Cultural Competence in Support of Indigenous-Led Evaluation: A Necessary Step toward Indigenous-Led Evaluation ». *Canadian Journal of Program Evaluation*, 34(2) : 368-384 (Extraits). Traduction par Carine Gazier et Thomas Delahais. Article originellement publié en *open access*.]

Je commence ma présentation aujourd'hui par un *mihimihi* (une introduction traditionnelle), comme je commence chaque présentation, que ce soit chez moi en *Aotearoa* Nouvelle-Zélande ou en tant que visiteuse sur la terre d'un autre peuple. Pour ce faire, j'identifie mes liens avec les lieux importants pour mon peuple, nos montagnes, nos rivières et nos *marae* (lieux de rassemblements traditionnels) et je reconnais mes relations avec les gens d'hier et ceux d'aujourd'hui. Je salue également les gardiens et gardiennes traditionnel-le-s de cette terre, attestant ainsi de mon statut de visiteuse.

Mon *mihimihi* affirme mon *whakapapa* (généalogie et identité) et mes liens avec la terre, l'environnement naturel, les ancêtres, les esprits et le cosmos (Marsden et Royal, 2003). Il nous rappelle notre lien avec les gens et le lieu (Henare, 1988; LaFrance et Nichols 2010), que nous sommes toute-s connectés à une source spirituelle et que nous vivons dans un réseau de réciprocité les un-e-s avec les autres et avec toute la création (Spiller et Stockdale, 2012).

L'entrée dans les communautés *Māori*, comme dans de nombreuses communautés autochtones, implique généralement d'autres rituels de rencontre tels que le *pōwhiri* (accueil officiel), le *whaikōreoro* (discours) et le *karakia* (chants rituels ou prières). Collectivement, ces protocoles culturels marquent un mouvement depuis le monde contemporain dominé principalement par les pratiques occidentales vers un monde *Māori* où les valeurs et les normes *Māori* dominent; et ils nous aident à nous préparer au contexte dans lequel nous allons travailler (Wehipeihana et McKegg, 2018).

Dans cette présentation, je définis l'évaluation autochtone, je décris la raison d'être de l'évaluation menée par les personnes autochtones, je propose une stratégie pour la soutenir et je partage un modèle (Wehipeihana, 2013) permettant aux évaluateurs et évaluatrices non autochtones de réfléchir à leur positionnement et à leurs méthodes de travail afin de révéler les dynamiques de pouvoir qui font obstacle à une évaluation autochtone menée par les peuples autochtones.

Définition de l'évaluation autochtone

Lorsque je pense à l'évaluation autochtone (ou dans mon cas à l'évaluation Kaupapa Māori), je pense à une évaluation par des personnes autochtones, pour les personnes autochtones, avec des autochtones et en tant qu'autochtones; et où il n'y a pas de rôle tenu pour les non-autochtones, sauf sur invitation. En Aotearoa Nouvelle-Zélande, « par les Māori, pour les Māori, avec les Māori » (Cram, 2016; Cram, Chilisa et Mertens, 2013) fait partie du programme de recherche et d'évaluation depuis les années 1980 (et plus récemment « en tant que Māori » : (Durie, 2001; Wehipeihana, McKegg et Pipi, 2015) (voir tableau 1).

L'autodétermination des peuples autochtones est fondamentale

L'évaluation Kaupapa Māori trouve ses fondements dans le principe du Tino Rangatiratanga (commandement principal) qui figure dans la version Māori du Traité de Waitangi signé par la Couronne britannique et les rangatira (chefs) en 1840. Tino Rangatiratanga a été interprété comme une traduction du terme « autodétermination » et fait référence à la détermination par les Māori des questions qui ont un impact sur les Māori. Tino Rangatiratanga a été au cœur des aspirations des Māori depuis la signature du traité en 1840 et le reste aujourd'hui.

Tableau 1. Déterminants fondamentaux de l'évaluation des *Maoris* et des peuples autochtones

Par les Maoris (par les personnes autochtones)	Il s'agit d'une évaluation conduite par des Maoris et dans laquelle les Maoris ont l'autorité et le pouvoir de prendre des décisions au sujet de la conception, des méthodes, des critères d'évaluation et des méthodes de travail de l'évaluation.
Pour les Maoris (pour les personnes autochtones)	Il s'agit de s'assurer que l'évaluation présente des avantages évidents pour les Maoris et que les aspirations des Maoris sont reconnues dans le cadre de l'évaluation.
Avec les Maoris (avec les personnes autochtones)	Où les Maoris constituent la majorité de l'équipe d'évaluation; l'évaluation tient compte des contextes tribaux et communautaires et observe et utilise avec respect le *te reo Māori me Ngā tikanga* (langue et pratiques culturelles *Māori*).
En tant que Maoris (en tant qu'autochtones)	Il s'agit d'une évaluation guidée, informée et étayée par le *kaupapa Tuku iho* (valeurs culturelles transmises par les ancêtres) et le *tikanga Māori* (pratiques culturelles *Māori*).
Rôle des non-Maoris (non-autochtones)	Lorsqu'il n'y a pas de rôle automatique ou supposé pour les non- Maoris dans l'équipe d'évaluation et que la participation des non-Maoris se fait sur invitation uniquement.

L'évaluation *Kaupapa Māori* s'inspire également de la théorie *Kaupapa Māori* (Smith, 1997) et de son utilisation comme outil stratégique et politique pour faire progresser les aspirations et les programmes *Māori*. À *Aotearoa*, la théorie *Kaupapa Māori* est la pierre angulaire de la recherche et de l'évaluation avec les *Māori*. Cram (2001) définit le *Kaupapa Māori* comme une voie *Māori*, soutenue par les philosophies, les valeurs et les principes *Māori* qui portent une orientation ou une vision inspirante pour le bien-être culturel, social et économique des Maoris.

Dans le même temps, « par les Maoris, pour les Maoris » a émergé dans le discours politique ainsi que dans les visions d'avenir de l'autodétermination autour des revendications du Traité de *Waitangi* et de la théorie *Kaupapa Māori*. « Par les Maoris, pour les Maoris » était une position culturelle et politique qui affirmait le droit des Maoris à mener leur propre développement, c'est-à-dire à s'autodéterminer. Les attentes des Maoris en ce qui concerne la conduite de la recherche et de l'évaluation ont commencé à être reflétées dans des directives éthiques spécifiques aux Maoris, telles que celles élaborées par le *Health Research Council of New Zealand* (2010) et par des organismes gouvernementaux tels que *Te Puni Kōkiri* (1999)– le ministère du Développement des Maoris – et le ministère du Développement social (2004). Toutefois, le rôle et la pertinence des chercheurs et des chercheuses *Pākehā*[1] dans la recherche impliquant les *Māori* ont fait l'objet d'un vif débat. D'une part, les Maoris soutenaient qu'il n'y avait pas de place pour les *Pākehā* dans la recherche sur les Maoris (Walker, 1990). D'autre part, Smith (1990) a identifié quatre modèles de recherche culturellement appropriés pour répondre aux besoins de recherche des Maoris. Ces modèles[2] de recherche n'étaient pas considérés sans mérite à l'époque, pour atténuer le manque de ressources et de capacités de recherche *Māori* (Cram, 1997).

Mener une évaluation dans le cadre Kaupapa Māori signifie que les évaluateurs et les évaluatrices non maoris entreprennent une évaluation à l'invitation de la communauté Māori et en partenariat avec celle-ci (Mertens, 2009). En ce qui concerne les évaluateurs et les évaluatrices non maoris, leur participation à l'évaluation Kaupapa Māori s'apparente à la notion d'invités (Harvey, 2003). Dès le départ, être invité signifie accepter d'être sur le « territoire » de quelqu'un-e d'autre et être disposé-

1. Néo-Zélandais d'origine européenne.
2. Le modèle Tiaki (mentor), dans lequel la recherche est guidée par des Maori faisant autorité; le Whangai (adoption), où les scientifiques deviennent l'un des whānau; le modèle de partage du pouvoir, dans lequel le ou la scientifique et la communauté participent conjointement à la recherche; et le modèle Empowering Outcomes, où les chercheurs et chercheuses fournissent des informations et des réponses à des questions ou à des sujets sur lesquels les Maori veulent s'informer.

e à s'en remettre aux protocoles de son hôte ou hôtesse. Dans ce contexte, les invité-e-s comprennent que leur rôle sera déterminé par les Maoris et, idéalement, iels[3] auront une conscience aiguë de leur position et de leur statut « extérieur » dans l'évaluation (Brayboy et Deyhle, 2000).

Ainsi, ma définition de l'évaluation autochtone trouve ses racines dans la lutte des Maoris pour leur autodétermination ainsi que l'exercice du contrôle et de la prise de décision sur leur vie et les choses qui comptent pour elles et eux. Toutefois, les Maoris ne sont pas les seul-e-s à revendiquer un programme d'autodétermination. Les peuples autochtones du monde entier plaident depuis longtemps en faveur de l'autodétermination, et nous constatons que l'importance de l'autodétermination autochtone se reflète dans les articles 3 et 4 de la Déclaration des Nations Unies sur les droits des peuples autochtones (adoptée par l'Assemblée générale le 13 septembre 2007) (voir tableau 2).

Tableau 2 : Déclaration des Nations Unies sur les droits des peuples autochtones : articles 3 et 4

Article 3 : Les peuples autochtones ont droit à l'autodétermination. En vertu de ce droit, ils déterminent librement leur statut politique et poursuivent librement leur développement économique, social et culturel.
Article 4 : Les peuples autochtones, dans l'exercice de leur droit à l'autodétermination, ont droit à l'autonomie ou à l'auto-administration pour les questions relatives à leurs affaires intérieures et locales, ainsi qu'aux moyens de financer leurs fonctions autonomes.

3. NdT : Nous utilisons « iels » au lieu de « ils et elles » dans cet article afin de rendre compte de la diversité des genres présente dans la culture Maori, incluant deux identités de genre en dehors de la binarité : whakawahine et whakatane. Voir : Journet, Nicolas. « Les traditions du troisième sexe », Martine Fournier éd., *Masculin-Féminin. Pluriel.* Éditions Sciences Humaines, 2014, p. 95-98. Et *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History* (ed. Gilbert Herdt).

Pourquoi une évaluation menée par les personnes autochtones?

Beaucoup d'encre a coulé sur la place de la recherche et de l'évaluation dans les traditions occidentales, y compris la domination de cette discipline intellectuelle (Bishop, 2005; Smith, 1999; Wehipeihana et McKegg, 2018); le rôle de l'impérialisme, du colonialisme et de la mondialisation dans l'élévation et la valorisation des traditions occidentales de la recherche (Chilisa, 2012; Cram, 2009); et la recherche et l'évaluation en tant qu'exercice du pouvoir et de contestation politique des connaissances et des ressources (Bishop, 2005; Te Awekotuku, 1991; Teariki, Spoonley, et Tomoana, 1992).

Si l'on souhaite que l'évaluation fasse une différence positive dans la vie des peuples autochtones et dans les choses qui comptent pour ces derniers, alors le discours occidental dominant doit être tenu à l'écart. Le contrôle autochtone est essentiel si l'on veut que les valeurs, les principes et les méthodes de travail autochtones dominent l'évaluation autochtone et si l'on veut créer un espace sûr pour que les communautés et les évaluateurs et les évaluatrices autochtones puissent être autochtones.

Lorsqu'une évaluation est menée par les personnes autochtones, celles-ci sont plus susceptibles de posséder le capital culturel pour :

- Faciliter l'engagement respectueux et le respect des protocoles culturels; iels savent ce qui est important pour que les relations et l'évaluation démarrent dans de bonnes conditions (Durie, 2001);
- Utiliser les connaissances, les méthodes et les façons de travailler autochtones dans le cadre d'une évaluation; iels peuvent fournir une « lecture » culturelle ou une évaluation de l'adéquation culturelle ou de la pertinence des méthodes et outils de collecte de données pour les peuples autochtones (Goodwin, Sauni, et Were, 2015);
- Faciliter la compréhension de ce que la valeur et le bien signifient pour les peuples autochtones; iels peuvent engager le processus

visant à tirer une signification et à analyser [les matériaux collectés] pour s'assurer que la richesse, la subtilité et la nuance du sens ne sont pas perdues dans la traduction et garantir la validité culturelle des conclusions évaluatives (Kirkhart, 2010);

- Lorsque j'écoute des conférenciers et conférencières autochtones du monde entier, et que je lis les rares, mais de plus en plus nombreux documents portant sur l'évaluation autochtone, le critère « par » les personnes autochtones ou mené par des personnes autochtones n'est pas prépondérant, et l'évaluation autochtone est confondue avec des méthodologies attentives à la culture. À mon avis, cela a pour effet de fournir tacitement une mission et/ou de suggérer par défaut qu'il est « acceptable » que des évaluateurs ou évaluatrices non autochtones mènent des évaluations avec des peuples autochtones.

En *Aotearoa* Nouvelle-Zélande, les aspirations des Maoris au contrôle de leur propre vie et à l'autodétermination signifient qu'il n'est pas acceptable pour les non- Maoris de diriger des évaluations avec des *whānau* (famille élargie), *hapū* (clan ou sous-tribu), *iwi* (tribu) ou des organisations *Māori*. Toutefois, ce n'est pas la norme en dehors d'*Aotearoa* Nouvelle-Zélande, où les évaluateurs et évaluatrices non autochtones dirigent la plupart du temps des projets d'évaluation avec et dans les communautés autochtones. Cela découle en grande partie des hypothèses non remises en question de la communauté internationale de l'évaluation selon lesquelles iels peuvent entreprendre des évaluations de grande qualité lorsqu'iels travaillent dans des communautés autochtones ou interculturelles, où la culture, le contexte et la ou les langues sont différents des leurs; recueillir de bonnes données; procéder à une analyse culturellement valide; et porter des jugements évaluatifs solides (Wehipeihana *et al.*, 2010). Cette hypothèse est de plus en plus remise en question, mais les évaluateurs/-trices autochtones sont peu nombreux-ses et ont une influence limitée; et le pouvoir de décision est généralement dévolu aux bailleurs et bailleuses de fonds et aux évaluateurs/-trices non autochtones.

Ma définition de l'évaluation autochtone fait de l'initiative autochtone un critère essentiel, sans que les peuples non autochtones ne jouent un rôle présumé ou automatique, sauf sur invitation. Elle reflète mes convictions personnelles et professionnelles sur ce qu'il faut faire pour effectuer une évaluation de qualité avec les peuples autochtones. Si l'on peut apprendre beaucoup de choses sur les connaissances culturelles, certaines ne peuvent être connues et révélées que de l'intérieur de la culture. Pour ces raisons, « il n'y a pas de substitut au capital culturel (compréhension, connaissances et intuition) qui provient de l'appartenance à la culture » (Wehipeihana *et al.*, 2010 : 188).

Soutenir l'évaluation conduite par les personnes autochtones

L'un des arguments que j'entends le plus souvent est qu'il n'y a pas assez d'évaluateurs/-trices autochtones pour que les évaluations menées avec les peuples autochtones soient dirigées par des personnes autochtones. Et je suis d'accord, il est prioritaire d'avoir plus d'évaluateurs/-trices autochtones. En s'appuyant sur l'expérience d'évaluation d'Aotearoa Nouvelle-Zélande, voici quelques éléments sur lesquels la Société canadienne d'évaluation[4] peut s'appuyer pour soutenir l'évaluation menée par les personnes autochtones :

- Soutenir le développement d'évaluateurs/-trices autochtones et de l'évaluation autochtone.
- Établir des partenariats avec les peuples autochtones pour guider l'évaluation.
- Accroître la compétence culturelle des évaluateurs/-trices non

4. NdT : L'article reprend une présentation faite par l'autrice en conférence plénière devant la Société canadienne d'évaluation en 2018.

autochtones.
- Promouvoir les valeurs et les principes autochtones dans la pratique de l'évaluation au Canada.
- Comprendre l'importance des relations lorsqu'on s'engage avec les peuples autochtones.
- Promouvoir la crédibilité et l'inclusion des valeurs, méthodes et critères autochtones.

Soutenir le développement des évaluateurs et évaluatrices autochtones et de l'évaluation autochtone

Des ateliers d'évaluation communautaires, dont certains s'adressent explicitement aux Maoris et sont financés par des organismes de promotion de la santé et de développement des Maoris, ont offert aux Maoris des voies alternatives pour entrer dans le domaine de l'évaluation et pour ouvrir un chemin à l'évaluation dans leurs communautés. Des mentors d'évaluation, tant *Māori* que non *Māori*, ont encouragé et aidé les Maoris à acquérir des qualifications universitaires en évaluation.

L'*Aotearoa New Zealand Evaluation Association* [Association d'évaluation d'Aotearoa Nouvelle-Zélande] (ANZEA) a aidé les évaluateurs/-trices *Māori* (et du Pacifique) en fournissant et en finançant un espace de rencontre dédié lors de sa conférence annuelle. Les évaluateurs/-trices *Māori* déterminent l'utilisation de ce temps et partagent la manière dont iels ont appliqué le *matauranga Māori* (systèmes de connaissances *Māori*), le *te ao Māori* (perspectives et visions du monde *Māori*), ainsi que les connaissances tribales et les protocoles culturels dans leur pratique de l'évaluation. Cela contribue à l'élaboration de pratiques et de principes d'évaluation *Kaupapa Māori* et au renforcement d'un réseau d'évaluation *Māori*.

Lors de l'élaboration des normes d'évaluation et des compétences des évaluateurs/-trices d'*Aotearoa* Nouvelle-Zélande, il y a eu de nombreuses conversations et un dialogue sur ce qui importe pour l'évaluation en Nouvelle-Zélande, et sur ce que signifient la qualité et la bonté. Les évaluateurs/-trices *Māori* ont décidé qu'une des façons de l'exprimer était de le faire selon leurs propres termes ; cela a conduit à la formation de *Ma te Rae* (l'association d'évaluation *Māori*) – la première organisation autochtone d'évaluation au monde. *Ma te Rae* participe activement à *EVAL Indigenous*, a son propre programme de travail de développement et siège sans difficulté aux côtés de l'ANZEA.

Alors qu'iels étaient moins de cinq en 1999, les évaluateurs/-trices *Māori* étaient environ 50 en 2018[5].

Établir des partenariats avec les peuples autochtones pour guider l'évaluation

Une connaissance culturelle approfondie est nécessaire pour que les évaluateurs/-trices puissent travailler en toute sécurité et dans le respect des communautés autochtones. Il ne s'agit pas d'une connaissance superficielle qui vous permettrait d'intégrer la communauté en observant quelques protocoles culturels. Il s'agit plutôt d'une connaissance culturelle qui vous permettra de naviguer respectueusement, de convenir de façons de travailler, y compris des méthodes et des approches appropriées, et d'aider les peuples autochtones à exprimer leurs opinions, leurs valeurs et leurs expériences.

5. Évaluateurs/-trices qui s'identifient comme Māori sur la liste des membres de l'association d'évaluation Ma te Rae Māori en novembre 2018.

Les évaluateurs non autochtones doivent trouver un-e conseiller-e ou un-e guide issu-e de la communauté autochtone, par exemple un-e aîné-e de la tribu ou un-e chef-fe, pour les aider et faciliter l'évaluation. Ces conseillers et conseillères culturels fournissent des renseignements historiques et contextuels, mettent les évaluateurs/-trices en contact avec la communauté et les guident dans le respect des protocoles culturels.

Un mot d'avertissement : Même si vous êtes *Māori* ou autochtone, vous ne pouvez pas supposer que vos connaissances s'appliquent à toutes les tribus et communautés.

Accroître la compétence culturelle des évaluateurs et évaluatrices non autochtones

Tout d'abord, faites vos devoirs avant d'entrer en contact avec les communautés autochtones. Renseignez-vous sur l'histoire et le contexte actuel (Ormond, Cram, et Carter, 2006). Les peuples autochtones en ont assez de raconter encore et encore leur histoire, leur récit. Lorsque vous souhaitez dialoguer avec les peuples autochtones, faites cette démarche de recherche : elle est la preuve d'une réelle volonté et d'un engagement sincère envers les peuples autochtones.

Deuxièmement, cherchez à élargir vos réseaux autochtones. Commencez par les gens que vous connaissez, et élargissez vos horizons. Cela peut se faire par le biais des arts, du sport, d'un projet ou d'une organisation communautaire. Il ne s'agira peut-être pas de personnes autochtones avec lesquelles vous collaborerez par la suite, mais vous développerez votre connaissance et votre compréhension « générale » de la culture autochtone, ce qui vous aidera dans vos futures actions.

Troisièmement, engagez-vous dans un développement professionnel pertinent. Il peut s'agir notamment d'ateliers d'évaluation spécifiquement destinés aux autochtones ainsi que d'ateliers de perfectionnement professionnel dans des domaines connexes, par exemple l'évaluation attentive à la culture, ou des ateliers portant sur l'équité, la diversité et la lutte contre le racisme. Vous pouvez également vous tourner vers des domaines autres que l'évaluation, tels que le développement communautaire, la justice sociale, les arts ou l'environnement – où l'on peut trouver des idées sur les façons de travailler avec les peuples autochtones, ainsi que des cours ou des documents d'études autochtones proposés par des établissements d'enseignement supérieur.

Mettre en exergue les valeurs et les principes autochtones dans la pratique de l'évaluation au Canada

En *Aotearoa Māori*, les valeurs et les principes ont été élevés dans la pratique de l'évaluation par l'association nationale. L'ANZEA a inclus les principes du Traité de *Waitangi* – partenariat, protection et participation – dans sa constitution. L'inclusion des principes du Traité établit le caractère unique de l'évaluation en *Aotearoa* et concentre notre avance en matière d'évaluation sur ce qui est nécessaire pour assurer l'inclusion et la participation des peuples autochtones et des perspectives autochtones dans tous les aspects de l'évaluation (Wehipeihana *et al.*, 2014). Ceci envoie également un message clair aux organisations qui financent l'évaluation quant à la centralité des valeurs et des principes *Māori* et leur intégration dans la pratique de l'évaluation en *Aotearoa*. Cela a été évident lors de l'élaboration des normes d'évaluation et des compétences des évaluateurs néo-zélandais et de l'intégration des perspectives *Māori* dans ces publications portant sur les pratiques professionnelles en *Aotearoa*.

Comprendre l'importance des relations avec les peuples autochtones

Dans les contextes *Māori* et autochtones, les relations sont le ciment, le point d'ancrage et la monnaie d'échange d'une collaboration efficace et respectueuse (Wilson, 2008) – et d'évaluations réussies. C'est à travers les relations que se déroule le processus d'évaluation. Et c'est dans les relations et la collaboration que la confiance relationnelle est bâtie. La confiance relationnelle ouvre la voie à des liens plus profonds et plus significatifs.

Les relations ne sont donc pas une question que l'évaluateur se contente d'examiner; elles sont inextricablement liées à la collaboration avec les peuples autochtones et, par conséquent, à l'évaluation autochtone (Wehipeihana *et al.*, 2015).

Plaider pour la crédibilité et l'inclusion des valeurs, méthodes et critères autochtones

Bien qu'il y ait eu des progrès au niveau gouvernemental, qui exige expressément l'utilisation ou la prise en compte des principes *Kaupapa Māori* dans l'évaluation impliquant les *Māori*, d'importants problèmes subsistent. Le rôle des *Māori* en tant que directeurs et directrices ou membres principales et principaux des équipes d'évaluation, et le respect des protocoles culturels sont relativement bien établis. Toutefois, il reste des défis à relever pour déterminer ce qui est considéré comme des preuves crédibles, des méthodes acceptées et l'inclusion des valeurs et principes culturels *Māori* dans l'élaboration des critères d'évaluation.

Le ministère de la Santé de la Nouvelle-Zélande a accepté et promu les cadres nationaux de santé *Māori* fondés sur la culture (ministère de la Santé, 2015 – *Te Whare Tapa Whā* (Les quatre piliers de la santé) (Durie, 1994), *Te Pae Mahutonga* (Constellation de la Croix-du Sud – quatre éléments de la promotion de la santé) (Durie, 1999) et *Te Wheke* (La pieuvre – définir la santé de la famille) (Pere, 1982), contribuant ainsi à l'élévation de l'importance de la *tikanga Māori* (valeurs culturelles *Māori*), en faisant ainsi un élément essentiel de l'évaluation des résultats de la santé des Maoris. En outre, en 2019, des signes positifs indiquent que les valeurs et principes *Māori* ont été incorporés dans les cadres nationaux de redevabilité pour tous les Néo-Zélandais, y compris les Maoris : « et que le bien-être considéré d'un point de vue autochtone déplace le discours de politiques publiques au-delà des constructions occidentales du bien-être et permet une meilleure expérience vécue du bien-être pour tous » (Te Puni Kōkiri and Treasury, 2019 : 1).

Les approches autochtones du cadre d'analyse des conditions de vie [*Living standards framework*] élaboré par *Te Puni Kōkiri* (Ministère du développement *Māori*) et le Trésor, bien qu'il soit axé sur le bien-être des *Māori* en particulier, présente une façon d'envisager le bien-être, qui peut être appliquée à l'ensemble des populations d'*Aotearoa* Nouvelle-Zélande (voir encadré 1). Le cadre offre un moyen de rendre compte des diverses valeurs et croyances qui stimulent le bien-être des gens et positionne le secteur public néo-zélandais dans la bonne voie pour faire progresser le bien-être d'une manière différente, en cherchant à répondre aux différents besoins, intérêts et aspirations de tous les Néo-Zélandais (Te Puni Kōkiri and Treasury, 2019).

Encadré 1. Une approche autochtone du cadre d'analyse des conditions de vie

Il n'y a pas une seule façon d'envisager le bien-être. Les gens voient le bien-être différemment selon leurs valeurs, leurs croyances et leurs normes sociales. La façon dont les Maori perçoivent le bien-être est différente de celle des autres Néo-Zélandais-e-s. Elle s'inspire de la *te ao Māori* (vision du monde Maori), dans laquelle, par exemple, la *whenua* (terre) n'est pas considérée uniquement pour son potentiel économique, mais à travers les liens familiaux et spirituels définis par des concepts culturels tels que le *whakapapa* (généalogie) et *kaitiakitanga* (intendance). La perspective *te ao Māori* du bien-être est également influencée par des expériences de vie – semblables à celles d'autres populations autochtones du monde entier –, d'importantes disparités et d'un accès inéquitable aux outils, aux ressources et aux possibilités qui constituent le fondement du bien-être. *Te Tiriti o Waitangi*, document fondateur d'Aotearoa Nouvelle-Zélande, accorde une grande importance au partenariat, à la protection active des intérêts *Māori* et à la réparation des torts passés – y compris la disparité et les inégalités persistantes dont souffrent les Maori et leur capacité à accéder et à bénéficier des différents capitaux sous diverses formes. Pris ensemble, ils transmettent une obligation pour la Couronne et les Maori de travailler ensemble. Pour ce faire, la Couronne – ministres, départements et autres organismes – doit chercher à comprendre le *te ao Māori*, en particulier en ce qui concerne l'amélioration du bien-être des *whānau* aujourd'hui et pour les générations à venir. Heureusement, le *te ao Māori* offre une façon de considérer le bien-être à travers un système holistique, robuste et existant de longue date. (*Te Puni Kōkiri* & Treasury, 2019)

Les Maoris défendent depuis longtemps que ce qui est bon pour les pour tou-te-s les Néo-Zélandais-es, et les approches autochtones du cadre d'analyse des conditions de vie sont un exemple de principes et de perspectives fondés sur la culture et intégrés dans un cadre d'évaluation national. Le message clé pour les évaluateurs/-trices est qu'il faut promouvoir et défendre l'inclusion des valeurs autochtones dans l'évaluation avec les peuples autochtones, et en faire valoir les avantages pour les peuples non autochtones. Il sera également important de persévérer dans ces efforts, car d'après l'expérience néo-zélandaise, la remise en question et le changement des valeurs profondément enracinées ne se font pas en un jour et ne s'obtiennent généralement pas par des réformes globales et radicales. D'après notre expérience, le changement des valeurs fondamentales passe par un radicalisme intransigeant et progressif.

Un changement de paradigme est nécessaire

Les évaluateurs et les évaluatrices non autochtones occupent une position privilégiée qui leur confère l'autorité et le pouvoir de définir la réalité, de porter des jugements influents sur les autres et de faire en sorte que ces jugements soient considérés comme exacts et valides (Johnson, 2001; Kirkhart, 2015; Sanakar, 2017). Il peut donc être difficile pour les évaluateurs/-trices non autochtones de changer leur pratique et de renoncer au pouvoir et aux privilèges. Iels doivent vouloir faire les choses différemment ou avoir une raison de voir le monde d'un autre œil. Cela peut être pour des raisons de justice sociale (Greene, 2011; Mertens, 2009), en raison d'une éthique de la sollicitude (*care*), ou en suivant le principe d'Agir sans nuire (*Do not harm*), ou alors iels peuvent être convaincu-e-s par des arguments méthodologiques ou pratiques tels que la validité multiculturelle (Kirkhart, 2010). Quelle que soit leur motivation, il ne s'agit pas simplement de ce qu'iels savent et comment ils font des évaluations; il s'agit fondamentalement de la façon dont iels perçoivent le monde – et donc un ou plusieurs changements de paradigme seront nécessaires. [Ils sont listés ci-dessous.]

… d'une évaluation transactionnelle à une évaluation relationnelle

L'évaluation transactionnelle consiste à mettre l'accent sur les processus d'évaluation, les méthodes, les outils et les délais nécessaires pour entreprendre une évaluation. Greene (2005) nous rappelle que l'évaluation porte non seulement sur ce que nous faisons, mais aussi sur notre identité et notre position par rapport aux autres. Par conséquent, le changement de paradigme vers une évaluation relationnelle reconnaît

que les relations sont la principale façon dont les peuples autochtones s'impliquent (Wehipeihana, 2013) et considère l'évaluation comme intrinsèquement relationnelle.

… de l'évaluateur/-trice aux peuples autochtones en tant qu'expert-e-s

L'évaluateur/-trice en tant qu'expert-e considère les évaluateurs/-trices comme des personnes qui possèdent généralement des qualifications, une expertise et une expérience en matière d'évaluation typiquement occidentales (Wehipeihana et McKegg, 2018). Le changement de paradigme vers les peuples autochtones en tant qu'expert-e-s affirme les connaissances culturelles uniques des peuples autochtones pour entrer, naviguer et dialoguer avec leur peuple (Durie, 2001; Smith, 1999) et démontre que la valeur des connaissances et de l'expérience occidentales diminue dans les contextes autochtones (Wehipeihana, 2013).

… de l'évaluation en tant que processus indépendant à l'évaluation en tant que processus connecté

L'évaluation, en tant que processus indépendant et impartial est une manière de privilégier les méthodes occidentales ou l'imposition de frontières relationnelles sur lesquelles se basent la validité et la crédibilité des jugements en matière d'évaluation. Le changement de paradigme en faveur d'une évaluation connectée aux autochtones et aux communautés, est essentiel à l'exactitude, à la crédibilité et à la validité culturelle des jugements évaluatifs (Wehipeihana et McKegg, 2018).

… de l'évaluation menée par des évaluateurs/-trices non autochtones à l'évaluation menée par des évaluateurs/-trices autochtones

Lorsque les évaluations dans les communautés autochtones sont gérées par des évaluateurs/-trices non autochtones, les évaluateurs/-trices autochtones ne sont pas toujours présent-e-s lorsque des documents clés sont élaborés et acceptés, ni lorsque des décisions importantes sont prises. Même lorsque des évaluateurs/-trices autochtones sont présent-e-s, iels sont souvent moins nombreux-ses, ont un statut inférieur aux autres, ou évoluent dans des environnements dans lesquels il n'y a pas de volonté d'envisager des options ou des points de vue alternatifs. La question sous-jacente est donc celle du pouvoir et du contrôle sur la prise de décision dans l'évaluation (Wehipeihana *et al.*, 2010). Lorsque les évaluations sont gérées par des évaluateurs/-trices autochtones, iels décident (la plupart du temps) de ce qui est important et donc prioritaire, de la façon dont les ressources sont allouées, des critères de jugement, de ce qui est considéré comme une preuve valide et de la façon dont les résultats sont rapportés.

Vers un contrôle autochtone de l'évaluation autochtone

Tous les changements de paradigme impliquent de concéder le contrôle ou, au minimum, de partager le pouvoir et l'autorité avec les peuples autochtones. J'ai mis au point un modèle[6] (Wehipeihana, 2013) pour déterminer dans quelle mesure les évaluateurs/-trices partagent le pouvoir et la prise de décision en évaluation avec les peuples autochtones.

Chaque aspect du modèle est brièvement abordé dans le tableau 3. Le modèle invite les évaluateurs/-trices à réfléchir à leurs pratiques en matière d'évaluation : c'est la première étape vers une plus grande part des peuples autochtones à la prise de décision dans les évaluations, dans le but clair de contribuer à une évaluation menée par les personnes autochtones. Lorsqu'il est utilisé comme un outil d'auto-évaluation, il indique comment le pouvoir est partagé et dans quelle mesure les peuples autochtones exercent un contrôle sur la prise de décision dans le cadre de l'évaluation.

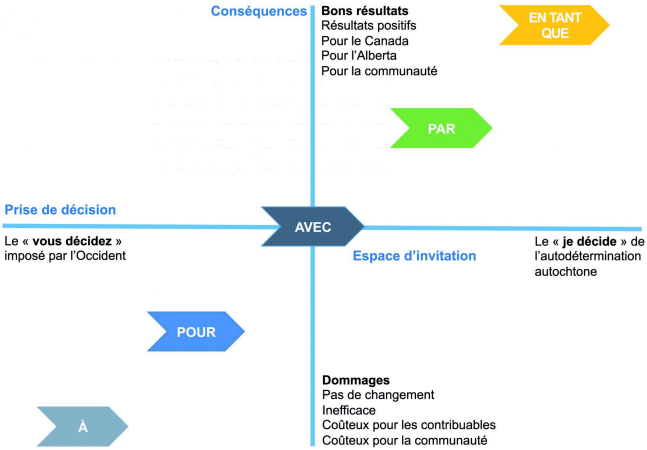
Quelques questions de réflexion initiales ont été élaborées pour soutenir un processus d'auto-évaluation :

- Qui est à l'origine de l'évaluation et qui en bénéficie?
- Quelle différence fera-t-elle pour les peuples autochtones?
- Qui la réalisera, et cette personne convient-elle à la communauté?
- Comment l'évaluation sera-t-elle réalisée, et la méthodologie est-elle « adaptée » à la communauté?

6. Le modèle était initialement intitulé "Un cadre pour renforcer le contrôle par les communautés autochtones" pour la conférence AES 2013 Wehipeihana (2013). Peu après (et aussi en 2013), je l'ai appelé "Localiser la pratique de l'évaluation : l'évaluation comme expression du pouvoir, du contrôle et des conséquences". Pour l'exposé liminaire de la CES 2018, je l'ai appelé "Cadre d'évaluation des progrès des évaluateurs vers la compétence culturelle autochtone et l'évaluation dirigée par les autochtones". J'ai tendance à changer le titre pour transmettre un message clé ou un objectif en fonction du public ou du contexte.

- Quel est (ou sera) le rôle des populations ou communautés autochtones dans l'évaluation? Ce rôle est-il doté de ressources? Que pensent-iels de leur rôle dans l'évaluation? Correspond-il à leurs attentes? Et si ce n'est pas le cas, peut-il être modifié?
- Quelles sont les possibilités pour vous ou pour l'évaluation de faire évoluer le statu quo?
- Que faut-il faire pour soutenir ce mouvement?

Figure 1 : Le modèle de Wehipeihana de 2013



Toutefois, des questions réflexives en chaussant des lunettes autochtones – peuvent être posées à n'importe quelle étape du processus d'évaluation, du démarrage de l'évaluation (par exemple, qui a entrepris l'évaluation et la communauté autochtone a-t-elle été consultée?); à la gestion de l'évaluation (par exemple, qui est partie prenante, les

autochtones sont-iels représenté-e-s et les processus décisionnels sont-ils équitables?); en passant par la conception de l'évaluation (par exemple, les principes et méthodes autochtones sont-ils inclus dans la conception de l'évaluation et les évaluateurs/-trices autochtones, et idéalement les populations locales, font-iels partie-s de l'équipe d'évaluation?), par l'analyse des données (par exemple, existe-t-il un processus pour vérifier l'exactitude et la validité culturelle de l'analyse des données et des conclusions de l'évaluation?); et enfin par la diffusion (par exemple, comment les conclusions de l'évaluation seront-elles communiquées aux communautés autochtones?).

Conclusion

Cet article a plaidé pour que les personnes autochtones soient au centre de l'évaluation autochtone. Il plaide en faveur d'un changement de paradigme personnel de la part des évaluateurs et des évaluatrices non autochtones et des financeurs et des financeuses des évaluations, car il est nécessaire de bouleverser leurs hypothèses prises pour acquises [selon lesquelles iels doivent avoir le] contrôle, et de modifier radicalement l'équilibre des pouvoirs en plaçant le contrôle entre les mains des peuples autochtones. Décoloniser sa pratique de l'évaluation est une expérience profondément personnelle. Cela exige une introspection, un examen approfondi de ses valeurs et de ses croyances, un examen de conscience, la mise en évidence et l'interrogation d'hypothèses et de préjugés implicites. Il faut également faire preuve d'humilité, s'ouvrir à d'autres perspectives et visions du monde, reconnaître les privilèges non mérités, partager le pouvoir et, en fin de compte, s'en défaire. Enfin, il faut du courage pour s'engager dans un voyage éprouvant à la fois sur le plan personnel et professionnel. Cela requiert donc aussi de la ténacité, afin de tenir le cap et résister aux cyniques, aux critiques et aux racistes. C'est un premier pas important.

Dans le même temps, il faut s'attaquer aux obstacles institutionnels et structurels en soutenant le développement d'évaluateurs/-trices autochtones et d'une communauté autochtone d'évaluation; en élevant les valeurs et les principes autochtones dans les pratiques canadiennes de l'évaluation; en plaidant pour la crédibilité et l'inclusion des valeurs, méthodes et critères autochtones dans l'évaluation; et en construisant la compétence culturelle des évaluateurs/-trices non autochtones et des bailleurs et bailleuses de fonds.

Tableau 3. Description des éléments du modèle Wehipeihana (2013)

Prise de décision	L'axe horizontal est un continuum de prise de décision et de contrôle passant du contrôle occidental (vous décidez) au contrôle autochtone (je décide).
Conséquences	L'axe vertical est un continuum de conséquences passant de bons résultats et de résultats positifs, à aucun changement, à des résultats inefficaces ou à des dommages.
Évaluation de	Les peuples autochtones n'ont pas leur mot à dire et n'ont pas de contrôle sur l'évaluation et il n'y a pas de bénéfices clairs pour elleux. L'évaluation est faite au sujet des peuples autochtones, et la vision occidentale du monde ainsi que la science occidentale prévalent.
Évaluation réalisée pour	S'efforce d'améliorer la situation des peuples autochtones, mais la consultation des peuples autochtones est minimale. L'évaluation est conçue et gérée sans tenir compte des valeurs, des principes et des priorités autochtones, et le pouvoir décisionnel appartient à l'évaluateur/-trice. L'évaluation est paternaliste, et les visions du monde et la science occidentales prévalent.
Évaluation réalisée avec	Le pouvoir et la prise de décisions sont partagés et négociés. C'est un espace de collaboration et de participation. Les approches et les visions du monde autochtones et occidentaux sont utilisées.
Évaluation réalisée par	Les peuples autochtones ont le contrôle de l'évaluation, et les méthodes ainsi que la vision du monde autochtones prévalent. Les évaluations peuvent utiliser des méthodes occidentales.
Évaluation réalisée en tant que	Les peuples autochtones ont le contrôle de l'évaluation, et le savoir et la science autochtone sont la norme. La légitimité et la validité des principes et des valeurs autochtones sont considérées comme acquises. Elle n'exclut pas les méthodes occidentales, mais ne les inclut que dans la mesure où elles sont jugées utiles.
Espace d'invitation	Le concept de l'espace d'invitation est celui où l'évaluation est contrôlée par les peuples autochtones, et les évaluateurs/-trices non autochtones acceptent que leur participation se fasse sur invitation uniquement.

Bibliographie

Bishop, Russell. 2005. « Freeing ourselves from neocolonial domination in research: A Kaupapa Māori approach to creating knowledge ». in *The SAGE handbook of qualitative research*, édité par N. K. Denzin et Y. S. Lincoln. Thousand Oaks: Sage Publications, p. 109-38.

Brayboy, Bryan M., et Donna Deyhle. 2000. « Insider-outsider: Researchers in American Indian communities ». *Theory into Practice* 29(3) : 163-69. doi : https://doi.org/10.1207/s15430421tip3903_7.

Chilisa, Bagele. 2012. *Indigenous methodologies*. Thousand Oaks: Sage Publications.

Cram, Fiona. 1997. « Developing partnerships in research: Pākehā researchers and Māori research ». *SITES* 35 : 44-63.

Cram, Fiona. 2001. « Rangahau Māori: Tona Tika, Tona Pono—The validity and integrity of Māori ». in *Research ethics in Aotearoa New Zealand*, édité par M. Tolich. Auckland, New Zealand: Pearson Education, p. 35-52.

Cram, Fiona. 2009. « Maintaining indigenous voices ». in *The SAGE handbook of social science research ethics*, édité par D. M. Mertens et P. E. Ginsberg. Thousand Oaks: Sage Publications, p. 308-22.

Cram, Fiona. 2016. « Lessons on decolonizing evaluation from Kaupapa Māori evaluation ». *Canadian Journal of Program Evaluation* 30(2) : 296-312.

Cram, Fiona, Bagele Chilisa et Donna M. Mertens. 2013. « The journey begins ». in *Indigenous pathways into social research: Voices of a new generation*, édité par D. M. Mertens, B. Chilisa, et F. Cram. Walnut Creek, CA: Left Coast Press, p. 11-40.

Durie, Mason. 1994. *Whaiora: Māori health development*. Auckland, New Zealand: Oxford University Press.

Durie, Mason. 1999. « Te Pae Mahutonga: A model for Māori health promotion ». *Health Promotion Forum of New Zealand Newsletter* 49 : 2-5.

Durie, Mason. 2001.« A framework for considering Māori educational advancement: The Hui Taumata Matauranga », Taupo, New Zealand: Massey University.

Goodwin, Debbie, Pale Sauni et Louise Were. 2015. « Cultural fit: An important criterion for effective interventions and evaluation work ». *Evaluation Matters—He Take Tō Te Aromatawai* 1 : 25-46.

Greene, Jennifer C. 2005. « Evaluators as stewards of the public good ». in *The role of culture and cultural context in evaluation: A mandate for inclusion, the discovery of truth and understanding*, édité par S. Hood, H. Frierson et R. Hopson. Charlotte, NC: Information Age Publishing, p. 7-20.

Greene, Jennifer C. 2011. « Evaluation and the public good: A mixed methods perspective ». Wellington, New Zealand.

Harvey, Graham. 2003. « Guesthood as ethical decolonising research method ». *Numen: International Review for the History of Religions* 50(2) : 125-46.

Health Research Council of New Zealand. 2010. « Guidelines for research on health research involving Māori ».

Henare, Manuka. 1988. « Ngā Tikanga Me Ngā Ritenga O Te Ao Māori. Standards and foundations of Māori society ». *The April report: Report of the Royal Commission on Social Policy Future (Wellington, New Zealand: The Royal Commission on Social Policy)* 3(1) : 3-42.

Johnson, Allan G. 2001. *Privilege, power and difference*. Mountain View, CA: Mayfield.

Kirkhart, Karen E. 2010. « Eyes on the prize: Multicultural validity and evaluation theory ». *American Journal of Evaluation* 31(3) : 400-413.

Kirkhart, Karen E. 2015. « Unpacking the evaluator's toolbox: Observations on evaluation, privilege, equity and justice ». *Evaluation Matters—He Take Tō Te Aromatawai* 1 : 7-24.

LaFrance, Joan, et Richard Nichols. 2010. « Reframing evaluation: Defining an indigenous evaluation framework ». *Canadian Journal of Program Evaluation* 23(2) : 3-31.

Marsden, Maori, et Te Ahukaramu Charles Royal. 2003. *The woven universe: Selected writings of Rev. Maori Marsden*. Otaki, N.Z.: Estate of Rev. Māori Marsden.

Mertens, Donna M. 2009. *Transformative research and evaluation*. New York: Guilford Press.

Ministry for Social Development. 2004. *Ngā Ara Tohutohu Rangahau Māori—Guidelines for Research and Evaluation with Māori*. Wellington, New Zealand: Centre for Social Research and Evaluation Te Pokapü Rangahau Arotake Hapori.

Ormond, Adreanne, Fiona Cram et Lyn Carter. 2006. « Researching our relations : Reflections on ethics and marginalisation ». *AlterNative: An International Journal of Indigenous Peoples* 2 : 174-93. doi : https://doi.org/10.1177/117718010600200108.

Pere, Rangimarie R. 1982. *Ako: Concepts and Learning in the Maori tradition*. Hamilton, New Zealand: University of Waikato Press.

Sanakar, M. 2017. *A stocktake on culturally responsive evaluation in and outside the UN system*. UNEG Strategic Objective 3: Evaluation informs UN system-wide initiatives and emerging demands. United Nations Evaluation Group, UNESCO Evaluation Office.

Smith, Graham H. 1990. *Research issues related to Māori education*. NZARE Special Interests Conference.

Smith, L. T. 1997. « The development of Kaupapa Māori: Theory and praxis ». Unpublished doctoral thesis, University of Auckland, Auckland, New Zealand.

Smith, Linda T. 1999. *Decolonising methodologies: Research and Indigenous peoples*. Dunedin, New Zealand: University of Otago Press.

Spiller, Chellie, et Monica Stockdale. 2012. « Managing and leading from a Māori perspective: Bringing new life and energy to organizations ». in *Handbook for faith and spirituality in the workplace*, édité par J. Neal. New York: Springer Publishing Company.

Te Awekotuku, Ngahuia. 1991. « He tikanga whakaaro: Research ethics in Māori community ». Discussion Paper, Wellington, New Zealand.

Te Puni Kōkiri. 1999. *Evaluation for Māori: Guidelines for government agencies*. Aroturuki me te Arotakenga Monitoring and Evaluation Branch. Wellington, New Zealand: Te Puni Kōkiri.

Te Puni Kōkiri and Treasury. 2019. *An Indigenous approach to the living standards frame work*. The Treasury Discussion Paper 19/01. Wellington, New Zealand: Treasury.

Teariki, Christine, Paul Spoonley, et Ngahiwi Tomoana. 1992. *Te Whakapakari te Mana Tangata: The Politics and Process of Research for Māori*. Palmerston North, New Zealand: Massey University.

Walker, Holly. 1990. *Māori attitudes on research*. Wellington, New Zealand: Department of Social Welfare.

Wehipeihana, Nan. 2013. A *vision for Indigenous evaluation*.

Wehipeihana, Nan, Robyn Bailey, E. Jane Davidson, et Kate McKegg. 2014. « Evaluator competencies: The Aotearoa New Zealand experience ». *Canadian Journal of Program Evaluation* 28(3) : 49-69.

Wehipeihana, Nan, E. Jane Davidson, Kate McKegg, et Vidhya Shanker. 2010. « What does it take to do evaluation in communities and cultural contexts other than our own? » *Journal of MultiDisciplinary Evaluation* 6(13) : 182-92.

Wehipeihana, Nan, et Kate McKegg. 2018. « Values and culture in evaluative thinking: Insights from Aotearoa New Zealand ». *New Directions for Evaluation* 158 : 93-107.

Wehipeihana, Nan, Kate McKegg et K. Pipi. 2015. « Cultural responsiveness through developmental evaluation: Indigenous innovations in sport and traditional Māori recreation ». in *Developmental evaluation exemplars: Principles in practice*, édité par M. Q. Patton, K. McKegg, et N. Wehipeihana. New York: Guilford Press, p. 25-44.

Wilson, Shawn. 2008. *Research is ceremony: Indigenous research methods*. Black Point: Fernwood.

Le regard de Thomas Archibald

THOMAS ARCHIBALD

Comme plusieurs ou même la plupart des évaluateurs et évaluatrices, mon entrée dans le domaine de l'évaluation était un concours de circonstances. Plutôt que de chercher explicitement à devenir évaluateur, je me compte parmi ceux qui pourraient être considérés comme un « évaluateur accidentel ». Fait intéressant, je n'ai jamais suivi qu'un seul cours universitaire sur l'évaluation ; un fait qui est assez ironique étant donné que j'enseigne maintenant l'évaluation aux étudiant-e-s en masters et en doctorat et mène des recherches sur l'évaluation. Mon parcours en évaluation, surtout mon entrée au hasard, est saillant pour ces réflexions sur le rôle des valeurs dans l'évaluation. Avant d'expliquer pourquoi, je dois tout d'abord féliciter les éditeurs et éditrices de cette anthologie pour avoir sélectionné un si excellent ensemble de textes représentatifs sur les valeurs en évaluation, ainsi que pour leur résumé merveilleusement clair, succinct et pertinent sur le sujet. Mes réflexions sur ce thème sont rendues plus faciles grâce à cet impressionnant travail éditorial.

En tant qu'évaluateur accidentel, en conséquence de mon manque de préparation formelle en évaluation, j'ai dû pratiquer pendant plusieurs années avant de constater ou de comprendre à quel point la logique de l'évaluation et, finalement, le rôle des valeurs étaient essentiels pour une évaluation de qualité. C'est quand enfin j'ai rencontré la définition classique et omniprésente de Scriven axée sur le mérite, l'intérêt et l'importance que j'ai commencé à saisir ces enjeux. Encore plus d'années se sont écoulées avant que j'apprenne ou réalise l'importance de la « logique de l'évaluation » discutée par Scriven, Fournier, Davidson et d'autres. La définition du mérite/de la valeur et la logique de l'évaluation sont des concepts clés pour aborder les processus par lesquels les évaluateurs/-trices parviennent à des jugements de valeur crédibles.

J'imagine que les évaluateurs/-trices émergent-e-s qui suivent un véritable cursus en évaluation rencontrent ces deux concepts au premier semestre, dans leurs cours sur les bases de l'évaluation.

Le fait que je n'aie appris la logique de l'évaluation qu'au bout de quelques années dans ma pratique est peut-être particulièrement ironique, car il s'agit d'un concept important par rapport à mon domaine d'intérêt principal, la notion de « pensée évaluative », ou la posture évaluative (Archibald, 2021; Archibald et Moussavou, 2016; voir la partie « Évaluatrice » pour en savoir plus sur ce concept). Cette ironie peut s'expliquer en partie par le fait que ma collègue Jane Buckley et moi-même en sommes venu-e-s à la notion de la pensée évaluative dans une perspective très pratique, enracinée en particulier dans la pratique du renforcement des capacités d'évaluation (RCE). Ainsi, nous avons concentré-e-s nos efforts sur les responsables de la mise en œuvre des programmes, pour les aider à mieux utiliser l'enquête évaluative pour améliorer leurs interventions, tout en répondant à leurs contraintes en matière de reddition de comptes. J'ai appris plus tard que ce type d'enquête, pas nécessairement conçu pour fournir des jugements sur la valeur, était peut-être un exemple de ce que Scriven (2016) a appelé la « phobie des valeurs » (*valuephobia*). Je me suis demandé si j'étais peut-être moi-même coupable de cette accusation.

Cette prise de conscience a suscité d'innombrables conversations avec Jane sur la question de savoir si et comment, dans notre approche du RCE et de la pensée évaluative, nous incluions l'étape du jugement de valeur assez clairement, de façon explicite. Pendant que ces conversations se poursuivaient, Jane et moi sommes parvenu-e-s à un consensus sur le fait que dans notre pratique d'évaluation (y compris notre pratique de facilitation du RCE), nous intégrons effectivement des processus de jugement de valeur. Cependant, ceux-ci se manifestent souvent dans des types de boucles de rétroaction plus courtes et plus rapides prises tout au long d'une évaluation, une sorte de pratique réflexive continue, plutôt qu'à la quatrième et dernière étape de la synthèse, comme d'autres pourraient le faire croire.

Cela étant dit, les textes présentés dans cette section font écho à mes réflexions et mes pratiques évaluatives des façons suivantes :

D'abord, l'ambiguïté persiste et certains résistent à l'idée de s'engager pleinement dans les questions de valeurs. Comme disent Hassall et ses collègues, dans un éditorial introduisant un numéro spécial de la *Revue d'évaluation d'Australasie* :

> « Au cours des cinquante dernières années, les hypothèses sur les valeurs comme subjectives, personnelles, émotionnelles et irrationnelles ont été contestées et réfutées, sur la base de recherches en psychologie sociale et en philosophie. Les philosophes reconnaissent maintenant que les faits, les valeurs, la subjectivité, la rationalité, l'émotion et l'objectivité sont empêtrés dans notre monde social et ne peuvent pas être facilement démêlés. En conséquence, l'idéal d'une science sociale sans valeur est [ou devrait être] largement rejeté ». (Hassall *et al.*, 2020 : 64)

Aussi, il faut réfléchir, surtout après avoir lu les textes présentés dans cette section de l'anthologie, aux risques de *ne pas* s'engager adéquatement sur ces questions du rôle des valeurs dans l'évaluation :

> « Alors, quel mal serait fait si les évaluateurs/-trices ignoraient les valeurs ? Dans un sens, aucun mal ne serait fait, car les évaluations contiendraient toujours des valeurs implicites. Les valeurs imprègnent inévitablement la sélection des variables indépendantes et dépendantes, le choix des questions et des parties prenantes, et le contexte social et politique d'où découlent de nombreuses évaluations. Les évaluateurs/-trices ne peuvent pas éviter les valeurs même s'ils essaient. Mais dans un autre sens, un véritable tort est fait si les évaluateurs/-trices traitent les valeurs pauvrement ou avec naïveté à travers leurs choix implicites ». (Shadish, 1994 : 35, cité par Hassall *et al.*, 2020)

Schwandt ne dit pas autre chose quand il affirme, prenant l'exemple de la définition du périmètre de l'évaluation, que la pensée évaluative est aussi un processus social dans lequel l'évaluateur/-trice a une responsabilité pour permettre aux parties prenantes d'exprimer leurs valeurs et leurs croyances, tout en mettant en évidence en quoi celles-ci affectent les choix qui sont faits (2018).

En réfléchissant à ce qui a été présenté ci-dessus, en ce qui concerne les conseils que j'aimerais donner à des évaluateurs et des évaluatrices émergent-e-s, je crois que certaines des orientations les plus importantes sur cette question sont les suivantes :

- Il faut apprendre et pratiquer la pensée évaluative, la pratique réflexive et la sagesse pratique dans votre travail d'évaluation, tout autant que les compétences techniques.
- Il faut être clair-e et explicite non seulement sur comment et pourquoi vous arrivez à certains jugements de valeur dans votre travail d'évaluation (le cas échéant), mais aussi sur ce que sont vos propres valeurs éthiques fondamentales. Cela peut vous aider à mieux comprendre votre rôle dans l'entreprise plus large qui utilise l'évaluation pour promouvoir le bien commun, que vous ayez ou non une position militante en faveur de valeurs sociales spécifiques, comme Khalil Bitar l'a demandé (voir Montrosse-Moorhead *et al.*, 2019), ou que vous souhaitiez engager une évaluation transformationnelle, à la manière de Jennifer Greene et de Donna Mertens.

Cette position, que certain-e-s peuvent considérer comme extrême, est bien présentée dans un nouveau livre de Schwandt et Gates, *Evaluating and Valuing in Social Research* (2021), qui propose un cadrage alternatif de l'évaluation en tant qu'activité qui développe délibérément la valeur plutôt que de simplement la déterminer. Surtout en cette ère de « post-vérité » et de « faits alternatifs », les évaluateurs et les évaluatrices aussi bien que

les chercheurs et chercheuses en sciences sociales ont la responsabilité primordiale de s'engager profondément et de manière réfléchie dans la façon dont ils et elles portent inévitablement des jugements de valeur.

Les textes de cette section de l'anthologie ainsi que le résumé habile des éditeurs et éditrices devraient contribuer grandement à aider tous les évaluateurs et les évaluatrices à mieux appréhender cette conversation en évolution sur le rôle des valeurs dans l'évaluation.

Bibliographie

Archibald, Thomas. 2021. « The Role of Evaluative Thinking in the Teaching of Evaluation ». *Canadian Journal of Program Evaluation* 35(3). doi : http://dx.doi.org/10.3138/cjpe.69753.

Archibald, Thomas, et Laurent O. Moussavou. 2016. « La pensée évaluative : Une activité mystérieuse et quotidienne ». *Éducation permanente* 208(3) : 33-40.

Hassall, Keryn, Amy M. Gullickson, Ayesha S. Boyce et Kelly Hannum. 2020. « Editorial ». *Evaluation Journal of Australasia* 20(2) : 63-67. doi : https://doi.org/10.1177/1035719X20931250.

Montrosse-Moorhead, Bianca, Khalil Bitar, Josette Arévalo, et Antonina Rishko-Porcescu. 2019. « Revolution in the making: evaluation "done well" in the era of the SDGs with a youth participatory approach ». *Evaluation for transformational change* 33.

Schwandt, Thomas A. 2018. « Evaluative thinking as a collaborative social practice: The case of boundary judgment making ». *New Directions for Evaluation* 2018(158) : 125-37. doi : https://doi.org/10.1002/ev.20318.

Schwandt, Thomas A., et Emily F. Gates. 2021. *Evaluating and valuing in social research*. 1re éd. New York: Guilford Press.

Scriven, Michael. 2016. « Roadblocks to recognition and revolution ». *American Journal of Evaluation*, 37(1) : 27-44. doi : https://doi.org/10.1177/1098214015617847.

Shadish, William R. 1994. « Need-based evaluation: Good evaluation and what you need to know to do it ». *Evaluation Practice* 15(3) : 347-58. doi : https://doi.org/10.1016/0886-1633(94)90029-9.

IV. L'ÉVALUATION EST-ELLE UNE SCIENCE?

Introduction : l'évaluation est-elle une science?

ANNE REVILLARD, THOMAS DELAHAIS, AGATHE DEVAUX-SPATARAKIS ET VALÉRY RIDDE

En quoi l'évaluation relève-t-elle de la science, et si c'est le cas, que lui apporte-t-elle? *Le Grand Robert de la langue française* définit la science comme un « ensemble de connaissances, d'études d'une valeur universelle, caractérisées par un objet et une méthode déterminés, et fondées sur des relations objectives vérifiables ». Bien que les définitions fassent débat, la démarche scientifique inclut généralement les notions de quête d'objectivité dans la production des savoirs (et à cet effet, l'utilisation systématique de méthodes en vue d'établir des faits sur la base de preuves empiriques), et d'aspiration à la généralisation (portée générale ou « universelle » des savoirs, plutôt que spécifique). L'évaluation s'est historiquement définie comme une pratique de science sociale appliquée, utilisant les méthodes de différentes sciences sociales, au service d'une analyse des enjeux et des conséquences des politiques publiques.

Ainsi entendue, elle relève essentiellement de la démarche scientifique par ses *méthodes*; son caractère « appliqué », en revanche, soulève des questions quant à sa capacité de généralisation : l'évaluation n'a-t-elle vocation qu'à répondre à des questions *ad hoc*, ou à produire des connaissances plus générales? La question de la montée en généralité est en réalité aussi présente en évaluation, que ce soit par des entrées méthodologiques (réflexion sur la « validité externe » des résultats d'une expérimentation par exemple), ou par une tentative de théorisation d'un apport spécifique de la démarche d'évaluation comme formation d'un jugement sur la valeur. Cet apport, défendu par Michael Scriven, justifie

selon lui de concevoir l'évaluation comme une « méta-discipline », dont la contribution est transversale par rapport aux autres disciplines scientifiques (Coryn et Hattie, 2007; Scriven, 1993).

Cette qualification disciplinaire révèle une autre dimension, plus institutionnelle, de la réflexion sur l'évaluation comme science. Au-delà de la caractérisation de la démarche (en quoi l'évaluation correspond-elle à ce que l'on peut attendre d'une science?), la discussion sur l'évaluation comme science révèle des enjeux relatifs à l'organisation de la production des savoirs : si l'évaluation est une science, doit-elle être considérée comme une discipline à part entière, bien distincte des autres, ou bien comme une pratique transversale? Indépendamment de cette dimension normative, quelle est la réalité de son inscription dans le champ universitaire de la production des savoirs?

À cet égard, le constat est celui d'une double marginalité : marginalité de l'évaluation dans le champ universitaire (du fait d'une difficulté à s'implanter comme trans-, méta- ou inter-discipline), mais aussi marginalité des acteurs et actrices universitaires dans le champ de l'évaluation. Comme le souligne Michael Patton par exemple, lorsqu'il établit un parallèle avec la médecine, en plus de l'évaluation comme science, celle-ci existe aussi comme technologie, comme pratique (Patton, 2018 : 12). Cette pratique concerne tout d'abord nombre d'actrices et d'acteurs en dehors du champ académique : consultantes et consultants dans le secteur privé, administrations, associations, etc. À cet égard, la question de savoir si l'évaluation est une science n'est pas un débat qui intéresse tou-te-s les évaluatrices et évaluateurs, mais en priorité celles et ceux du le monde académique, pour qui cet enjeu est important. Beaucoup d'autres seront plus attentives et attentifs à l'utilisation de l'évaluation qu'à son fondement scientifique (voir partie Utilité). Le monde académique, auquel s'intéresse cette partie, ne représente qu'un aspect d'un ensemble plus complexe des modalités d'institutionnalisation des pratiques d'évaluation (Jacob, Speer, et Furubo, 2015). En outre, les différents mondes de l'évaluation communiquent, ce qui contribue à enrichir l'évaluation comme science : comme nous

l'avons souligné au début de cette partie, l'évaluation a beaucoup pris appui sur les apports de différentes sciences sociales, et réciproquement, sa dimension appliquée et ancrée dans l'action a permis d'introduire en recherche des questionnements habituellement plus périphériques, notamment sur la place des valeurs et du jugement (voir partie Valeurs), et sur l'utilité des savoirs produits (voir partie Utilité).

Enfin, parallèlement aux enjeux épistémologiques et institutionnels, la revendication de l'évaluation comme science soulève aussi des enjeux d'opportunité et d'affichage. Le label scientifique peut ainsi être revendiqué ou mis à distance, en fonction des connotations qu'on lui associe. Dans le champ de l'évaluation, ce label est parfois utilisé, avec des connotations positives ou négatives, pour désigner l'évaluation fondée sur un certain type de méthode, les méthodes expérimentales, ou plus généralement les approches quantitatives. Dans les années 1990 en France par exemple, les tenantes et tenants de l'évaluation « pluraliste » tendaient ainsi à opposer la démarche d'implication des parties prenantes à des approches « positivistes » accusées de scientisme, dans une mise à distance de la science réduisant implicitement cette dernière a un certain type d'épistémologie et de méthodes (usage peu réflexif des méthodes quantitatives). La participation était alors mise en avant, contre la démarche consistant à fonder l'évaluation sur des méthodes scientifiques (Nioche, 2014). Plus récemment, alors que les méthodes expérimentales connaissent depuis quelques années un regain d'intérêt, on observe une tendance, chez certaines actrices et acteurs publics, à assimiler évaluation scientifique et évaluation quantitative, approche à laquelle on juge parfois nécessaire d'apporter un supplément d'âme en la combinant avec un dispositif consultatif ou participatif permettant de recueillir le point de vue des citoyennes et citoyens sur la politique étudiée.

Dans les deux cas, il en résulte une vision très appauvrie de la recherche évaluative, opposant le supposé scientisme des essais randomisés – là où Donald Campbell, Thomas Cook, Lee Cronbach ou William Shadish, par exemple, ont une conception beaucoup plus fine et mesurée de la portée des méthodes expérimentales (Cronbach, 1987; Shadish, Cook, et

Campbell, 2002), à une alternative de participation ou de consultation des parties prenantes en dehors de tout protocole scientifique, alors que les démarches de science participative montrent que la production scientifique et la participation peuvent en réalité être combinées de façon très fructueuse (Houllier et Merilhou-Goudard, 2016). Comme nous le verrons dans la partie consacrée aux approches paradigmatiques, la recherche évaluative, envisagée comme science, recouvre en réalité une diversité d'approches et de techniques d'investigation empirique (cf partie Paradigmes). La revendication de scientificité ne doit donc surtout pas être prise comme un choix méthodologique; elle englobe une grande variété de méthodes et de points de vue. Il importe toutefois de prendre acte des connotations possibles de la notion de « science » et des usages dont elle peut par conséquent faire l'objet dans le champ de l'évaluation.

Fondements : mobiliser la science pour l'évaluation

Comme nous le verrons au début de cette partie (**texte 1**, Shadish, Cook et Leviton), c'est dans le contexte de l'essor des politiques sociales sous la présidence Johnson aux États-Unis dans les années 1960 (« *Great society* »), notamment en réponse aux critiques conservatrices adressées à ces politiques, et sous l'effet d'une obligation légale d'évaluation, que se sont développées des évaluations prenant de plus en plus appui sur les méthodes de sciences sociales et revendiquant l'emploi d'une démarche scientifique. D'emblée, les compétences disciplinaires mobilisées sont diverses : psychologie, santé, éducation, sociologie, économie, etc. Par-delà cette diversité, ces évaluations ont en commun leur attachement à une démarche d'investigation empirique. Shadish *et al.* (1991) mettent en évidence la façon dont la massification de l'enseignement supérieur en sciences sociales est venue, à la même époque, rendre cette expertise disponible pour alimenter le vivier d'une nouvelle profession de l'évaluation. L'autrice et les auteurs montrent comment cette profession

s'est graduellement structurée avec ses revues, ses sociétés savantes, et la revendication graduelle d'une commune « logique de l'évaluation » (selon les termes de Scriven) par-delà l'extrême diversité des pratiques.

Le texte de Suchman **(texte 2)** illustre une des premières mises en forme de la conception de l'évaluation comme pratique scientifique, en distinguant la « recherche évaluative » de l'évaluation plus subjective du sens commun (le simple fait de porter un jugement sur quelque chose). On retrouve la notion de « recherche évaluative » chez de nombreuses autres figures fondatrices du champ de l'évaluation, par exemple Donald Campbell, Peter Rossi et Carol Weiss. Selon ces autrices et auteurs, la dimension scientifique de l'évaluation repose sur la systématicité des méthodes utilisées, empruntées à d'autres sciences sociales : pour elles et eux, c'est parce que l'évaluation mobilise ces méthodes qu'elle peut être considérée comme une science.

Simultanément, et comme le développe également Suchman, l'évaluation se distingue des autres sciences par son caractère *appliqué*, qui induit des contraintes et des préoccupations spécifiques. À la différence de la recherche fondamentale, la recherche évaluative (qu'il s'agisse par exemple de son questionnement, de son objet, des critères utilisés et des délais de réalisation) est soumise des contraintes externes au champ scientifique. La dimension appliquée de la recherche appelle aussi une préoccupation plus marquée vis-à-vis de l'utilité des savoirs produits : comme le souligne Suchman, l'évaluateur ou l'évaluatrice ne peut pas se satisfaire de l'idée selon laquelle « l'opération a été un succès même si le patient est décédé ».

Enfin, une plus forte imbrication de la recherche appliquée dans des logiques d'action induit, pour Suchman, un rapport différent, plus interventionniste, à l'objet de recherche. Alors que la recherche fondamentale se contente le plus souvent d'observer, de mesurer et de comprendre le réel, la recherche évaluative le manipule en modifiant les interventions existantes pour mettre en place des dispositifs de type expérimental permettant d'en tester l'efficacité. Bien qu'en pratique, les

recherches évaluatives ne se réduisent pas à des dispositifs de type expérimental, lesquels se sont par ailleurs diffusés au sein d'autres disciplines (par exemple la psychologie ou l'économie), la démarche d'évaluation a bien joué un rôle central dans l'essor de la pratique expérimentale, en rupture avec l'hypothèse partagée par différents courants fondateurs des sciences sociales définissant celles-ci comme des sciences de l'observation à la différence justement des sciences expérimentales impliquant une manipulation du réel – voir par exemple à ce sujet les réflexions de Durkheim sur l'utilisation de la comparaison comme substitut à l'expérimentation, dans une approche partageant par ailleurs la même épistémologie relative à l'imputation causale[1].

Le texte de Sandra Mathison **(texte 3)** sur la distinction entre recherche et évaluation permet de dresser un bilan des représentations couramment associées à cette vision de l'évaluation comme science appliquée plutôt que fondamentale (particularisation *vs* généralisation, recherche orientée vers l'action *vs* recherche comme fin en soi…), tout en soulignant la porosité de ces frontières (voir également sur ce point Levin-Rozalis, 2003). Mathison insiste par ailleurs sur l'apport méthodologique propre de l'évaluation, qui a été à l'origine de pratiques de recherche innovantes caractérisées par l'attention particulière accordée au point de vue des parties prenantes du programme étudié.

1. « Nous n'avons qu'un moyen de démontrer qu'un phénomène est cause d'un autre, c'est de comparer les cas où ils sont simultanément présents ou absents et de chercher si les variations qu'ils présentent dans ces différentes combinaisons de circonstances témoignent que l'un dépend de l'autre. Quand ils peuvent être artificiellement produits au gré de l'observateur, la méthode est l'expérimentation proprement dite. Quand, au contraire, la production des faits n'est pas à notre disposition et que nous ne pouvons que les rapprocher tels qu'ils se sont spontanément produits, la méthode que l'on emploie est celle de l'expérimentation indirecte ou méthode comparative. (…) Puisque (…) les phénomènes sociaux échappent évidemment à l'action de l'opérateur, la méthode comparative est la seule qui convienne à la sociologie. » (Durkheim, 2010).

Pour résumer, dans ce premier mouvement de théorisation de l'évaluation comme science, la revendication de scientificité de l'évaluation repose essentiellement sur les méthodes. Les modalités de développement de cette recherche évaluative portent en germe deux sources de fragilité pour l'institutionnalisation de l'évaluation en tant que discipline scientifique. La première est l'interdisciplinarité (la recherche évaluative puise dans des compétences disciplinaires très diverses) : épistémologiquement féconde, cette dernière est institutionnellement périlleuse dans un fonctionnement universitaire qui reste très marqué par une organisation en silos disciplinaires. D'autre part le caractère appliqué de l'évaluation, tout en étant source d'innovation théorique, induit une fragilité potentielle dans un contexte académique de valorisation de l'autonomie et du caractère autofinalisé de la science (la science comme fin en soi, et non comme moyen pour l'action).

Controverses : l'évaluation, une science à part?

Dans un texte datant de 1990, Gary Cox (**texte 4**) confirme cette difficulté d'institutionnalisation de l'évaluation dans le monde universitaire. Il note que l'évaluation à l'Université n'existe que comme pratique marginale de chercheurs et chercheuses dont la légitimité scientifique s'est construite sur d'autres bases (thématiques, méthodologiques). Il en résulte une difficulté à faire progresser les théories en évaluation dans la mesure où les dimensions institutionnelles et scientifiques sont en interaction. De fait, encore aujourd'hui, alors que l'évaluation dispose de ses revues, colloques et associations professionnelles (mêlant praticien-ne-s et chercheur-e-s), la recherche évaluative reste le plus souvent pratiquée dans le cadre (et souvent en marge) de disciplines universitaires instituées (éducation, santé, économie, urbanisme, etc.), plutôt que dans les départements dédiés à l'évaluation de programmes.

La difficulté, selon Scriven (**texte 5**), réside dans une méconnaissance ou un défaut de reconnaissance de l'apport propre de l'évaluation, apport qui justifie de la considérer comme une métadiscipline. Pour lui, cet apport ne réside pas tant dans l'appui sur des méthodes systématiques (qui diffèrent peu des méthodes utilisées par les différentes disciplines), mais bien plutôt dans une science du jugement, de la détermination de la valeur. Notant que les scientifiques ont tendance à mettre à distance la question des valeurs, il défend leur réintégration au cœur de l'activité scientifique, et l'apport fondamental de l'évaluation à cet égard. Cette réappropriation de la dimension des valeurs donne à la personne qui évalue un rôle plus actif que dans la représentation précédente d'une science sociale appliquée qui se contenterait d'appliquer les critères d'évaluation fixés par un commanditaire. L'évaluatrice ou l'évaluateur, selon Scriven, doit assumer une responsabilité dans le choix des critères d'évaluation, en mettant au premier plan non pas les demandes du ou de la commanditaire, mais les besoins et les droits des personnes visées par le programme évalué – responsabilité dont il souligne la dimension politique (cf partie Valeurs).

La caractérisation de l'évaluation comme métadiscipline amène également Scriven à insister sur la parenté de démarche entre l'évaluation de programme et les processus d'évaluation relevant de tout autre domaine, tel que l'évaluation de produits. À l'inverse de Shadish *et al.* (**texte 1**) qui insistent sur les spécificités de l'évaluation des programmes sociaux par comparaison avec les démarches d'évaluation pratiquées dans le secteur privé, Scriven met l'accent sur les similitudes de raisonnement d'un secteur à l'autre.

Perspectives : quelle(s) science(s) de l'évaluation pour demain?

Si elle a soulevé d'importants enjeux épistémologiques, la réflexion ouverte par Scriven sur l'évaluation comme métadiscipline n'en a pas moins eu des retombées institutionnelles très limitées dans le champ universitaire. En pratique, l'évaluation y existe très peu en tant que discipline séparée, et les chercheuses et chercheurs se revendiquant d'une démarche de recherche évaluative le font le plus souvent à partir de leurs disciplines d'appartenance. La question de l'ancrage et du développement institutionnels de l'évaluation dans le monde académique reste encore largement ouverte. Elle impose, en pratique, de réfléchir sur les modalités d'organisation d'une recherche qui reste interdisciplinaire bien plus que métadisciplinaire. C'est ce à quoi nous engage le texte de Steve Jacob (**texte 6**), invitant à l'hybridation des disciplines.

Dans un texte récent, Dana Wanzer apporte un utile complément à cette réflexion en schématisant les façons les plus courantes de penser les rapports entre recherche et évaluation, du *continuum* à l'imbrication mutuelle, en passant par l'idée d'un recoupement partiel **(texte 7**).

Enfin, à l'époque contemporaine, le questionnement sur l'évaluation comme science prend une tonalité particulière dans un contexte politique où la démarche scientifique, au sens le plus basique de l'établissement de faits objectifs, fait l'objet de vives attaques politiques. Dans un contexte marqué par le scientoscepticisme et l'essor des *fake news*, il devient plus important que jamais de réaffirmer l'intérêt d'ancrer l'évaluation dans une démarche d'investigation empirique systématique. C'est justement en marge d'une manifestation de défense de la science que Patton (**texte 8**) a proposé de parler de « science de l'évaluation » : « La science consiste à étudier de façon systématique comment le monde fonctionne. La science de l'évaluation consiste à étudier de façon systématique comment, et avec quel succès, des interventions visant à changer le monde fonctionnent » (Patton 2018 : 2). Patton met

simultanément en garde contre le risque de renforcer de ce fait une conception restrictive de la science (qui en pratique, se limiterait à l'expérimentation contrôlée de grande échelle). Défendre l'évaluation comme science, pour lui, c'est aussi défendre une conception pluraliste et multi-méthode de la pratique scientifique.

La réflexion sur l'évaluation comme science, et l'ensemble des textes traduits sont mis en perspective par Yves Gingras (UQAM), à l'aune de son expertise en histoire et sociologie des sciences, et notamment de son expérience de directeur scientifique de l'Observatoire des Sciences et des Technologies (OST) au Canada. Sa discussion incisive permet de poser à nouveau la question de la définition des sciences et du positionnement de l'évaluation. Faisant le lien entre les dimensions institutionnelles et épistémologiques de la réflexion que nous proposons, il souligne en quoi l'institutionnalisation de l'évaluation dans le champ universitaire favorise un discours de « scientifisation », parfois à distance de l'enjeu des usages de l'évaluation (cf partie Utilité).

Bibliographie

Coryn, Chris, et John Hattie. 2007. « The Transdisciplinary Model of Evaluation ». *Journal of Multidisciplinary Evaluation*.

Cronbach, Lee J. 1987. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass Publishers.

Durkheim, Émile. 2010. *Les règles de la méthode sociologique*. Paris : Flammarion – Champs classiques.

Houllier, François, et Jean-Baptiste Merilhou-Goudard. 2016. *Les sciences participatives en France : état des lieux, bonnes pratiques et recommandations*. Paris : MESRI.

Jacob, Steve, Sandra Speer et Jan Eric Furubo. 2015. « The institutionalization of evaluation matters: Updating the International Atlas of Evaluation 10 years later ». *Evaluation* 21(1) : 6-31. doi : 10.1177/1356389014564248.

Levin-Rozalis, Miri. 2003. « Evaluation and research, differences and similarities ». *The Canadian Journal of Evaluation* 18(2) : 1-31.

Nioche, Jean-Pierre. 2014. « L'évaluation des politiques publiques et la gestion en France : un rendez-vous manqué? » *Revue française de gestion* (8) : 71-84.

Patton, Michael Quinn. 2018. « Evaluation Science ». *American Journal of Evaluation* 39(2) : 183-200. doi : 10.1177/1098214018763121.

Scriven, Michael. 1993. « Hard-Won Lessons in Program Evaluation. » *New Directions for Program Evaluation* (58).

Shadish, William R., Thomas D. Cook et Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Wadsworth Publishing.

1. L'évaluation des programmes sociaux. Histoire, missions et théories

WILLIAM R. SHADISH, THOMAS D. COOK ET LAURA C. LEVITON

[Traduit de : Shadish, William R., Thomas D. Cook, Laura C. Leviton. 1991. *Foundations of Program Evaluation : Theories of Practice*. London: Sage. Excerpts from chapter 1 « Social program evaluation : its history, tasks, and theory », p. 19-35. Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Tout peut s'évaluer – y compris l'évaluation elle-même. Ce livre traite de l'évaluation des programmes sociaux. Ces programmes, et les politiques sociales qui les engendrent et les justifient, visent à améliorer le bien-être d'individus, d'organisations et de la société en général. Il est dès lors utile d'évaluer dans quelle mesure un programme social donné accroît effectivement le bien-être, comment il le fait et comment il peut le faire de manière plus efficace.

Une telle réflexion est urgente, car nous avons peu de critères clairs et partagés pour évaluer le mérite d'activités sociales. En quoi consiste « l'accroissement du bien-être »? Quelle que soit la façon dont on le définit, la définition n'est pas aussi partagée que ne l'est le critère du profit, utilisé pour évaluer les activités lucratives relevant du secteur privé. Même dans les rares programmes sociaux pour lesquels la plupart des parties prenantes s'accordent pour valoriser le même critère, ce critère n'aura jamais les mêmes propriétés pratiques que le profit – sa métrique simple en termes de dollars et de cents, ou sa plasticité lui permettant de dénombrer des résultats de nature aussi différente que le nombre de personnes employées, d'heures travaillées ou le volume de

produits expédiés. La plupart d'entre nous ont une perception intuitive de ce que l'argent « signifie », et de la valeur de différentes sommes d'argent. C'est moins souvent le cas concernant les résultats des programmes sociaux. La vision intuitive et la valeur comparée de différents volumes de cohésion familiale, de mobilité sociale, ou de frustration relative, ne sont pas aussi évidentes.

[…] Dans les vingt dernières années, les praticien-ne-s et les universitaires en évaluation de programmes se sont attelé-e-s à ces tâches et à d'autres qui leur sont liées. Dans un premier temps, ils ont commencé par emprunter des concepts et des pratiques à d'autres champs, notamment aux disciplines universitaires au sein desquelles ils et elles avaient été formé-e-s. Mais au fil que l'expérience en évaluation de programmes s'est accumulée, les évaluatrices et évaluateurs ont adapté ces concepts et ces méthodes, en ont inventé d'autres, et les ont combinées dans de nouvelles manières de pratiquer l'évaluation. De façon très nette, l'évaluation des programmes sociaux est un champ guidé par la pratique. […]

L'évaluation de programmes rejoint la médecine ou l'ingénierie dans l'accent qu'elle met sur la pratique. Pour autant, ces trois champs prennent aussi appui sur des théories qui n'ont pas de dimension immédiatement pratique. Les médecins apprennent les fondements de l'anatomie et de la physiologie, non pour y trouver des outils de pratique, mais pour comprendre les systèmes au sein desquels leur pratique se déploie […]. Les ingénieurs et ingénieures apprennent la physique, non pour disposer du mode d'emploi pour construire des fusées, mais parce que la physique leur donne les concepts dont ils et elles ont besoin pour comprendre et résoudre des problèmes dans le cadre de leur travail. On peut faire des observations similaires sur d'autres champs orientés vers la pratique. Ils tirent leurs théories fondamentales d'autres disciplines, mais utilisent ces savoirs et leurs propres expériences de résolution des problèmes pour développer des théories spécialisées adaptées spécifiquement aux exigences pratiques de leur travail. Ce serait une grosse erreur que de suggérer que la médecine ou l'ingénierie pourraient

fonctionner sans de telles théories. De la même façon, c'est une erreur que de négliger l'importance de la théorie en évaluation de programmes. [...]

Contexte et histoire de l'évaluation des programmes sociaux

[...] Nous allons nous intéresser aux théories et aux pratiques en évaluation depuis les années 1960, notamment concernant l'évaluation des politiques sociales. Ces efforts ont des origines intellectuelles dans des travaux antérieurs, notamment ceux de Tyler (1935) en éducation, de Lewin (1948) en psychologie sociale, et de Lazarslfeld (Lazarsfled et Rosenberg, 1955) en sociologie. Ils s'inscrivent aussi dans le contexte de la croissance économique rapide aux États-Unis après la Seconde Guerre mondiale, puis du rôle interventionniste pris par le gouvernement fédéral états-unien en matière de politiques sociales dans les années 1960, et enfin du nombre croissant de diplômé-e-s de sciences sociales intéressé-e-s par l'analyse des politiques publiques, les enquêtes quantitatives, les expériences de terrain, et l'ethnographie. Tous ces éléments ont posé le cadre de l'évaluation de programmes telle qu'elle se pratique aujourd'hui.

Les années 1960 et la « Great society »

L'évaluation des programmes sociaux telle qu'elle se pratique aujourd'hui a émergé dans les années 1960. Son développement résulte largement des politiques sociales initiées sous la présidence Kennedy et étendues sous les présidences Johnson et Nixon. Des programmes sociaux ont alors été lancés en matière d'éducation, de garantie du revenu, de logement, de santé et de droit pénal, principalement dans l'espoir de protéger les

Américains des effets négatifs de la pauvreté. La plupart de ces programmes ont été déployés très rapidement, en engageant d'énormes budgets, et ont été accompagnés de grandes attentes. [...]

Ces investissements considérables ont suscité d'importantes questions. Le Parlement se préoccupe du fait que les bénéficiaires de fonds fédéraux rendent compte de la façon dont ils les dépensent, en se souciant notamment de pouvoir anticiper l'évolution des dépenses et empêcher le versement de prestations indues. Mais le Parlement aspire aussi à ce que les programmes produisent certains des effets attendus tout en évitant le plus possible les effets secondaires négatifs. Jusqu'aux programmes sociaux des années 1960, ces fonctions parlementaires existaient plus sur le papier qu'en pratique. La croissance rapide du financement fédéral des programmes sociaux, mais aussi la couverture médiatique de cas de fraude, d'abus et de mauvaise gestion, ainsi que les réticences courantes envers la légitimité des programmes sociaux, ont poussé plusieurs parlementaires à investir ces fonctions de contrôle et de surveillance. Ces préoccupations ont augmenté au fil du temps, au fil de l'accroissement des budgets de défense, du taux d'inflation, et du déficit. Dans ce contexte, les défenseurs et défenseuses des programmes sociaux ont été de plus en plus appelé-e-s à démontrer que les fonds liés à chaque programme avaient bien été dépensés comme prévu, et de façon à produire des résultats souhaitables.

Mais le développement de l'évaluation a aussi été encouragé par un autre type de préoccupation politique. Certain-e-s observateurs et observatrices estimaient que les projets étaient mis en œuvre, au niveau local, selon des modalités ne correspondant pas aux intentions du gouvernement fédéral (Cronbach *et al.*, 1980; Cumming, 1976; House, 1980). Les parlementaires et agent-e-s de l'Exécutif au niveau fédéral voulaient se donner plus de prise sur ces projets, afin d'en assurer un meilleur contrôle. Inversement, d'autres percevaient ces nouvelles initiatives fédérales comme une menace sur l'autonomie du niveau local; leur objectif était de documenter les effets bénéfiques de projets

contrôlés localement (Feeley et Sarat, 1980). Dans les deux cas, un rôle important était assigné à l'évaluation, que ce soit pour déplorer l'existence de variations locales ou à l'inverse pour en faire l'éloge.

Des enjeux de gestion ont aussi nourri l'essor de l'évaluation. L'investissement fédéral massif dans les politiques sociales était un phénomène nouveau; peu de gestionnaires de ces nouveaux programmes avaient une expérience préalable en la matière. Le simple fait de lister les activités effectivement mises en œuvre dans le cadre d'un nouveau programme social donné était déjà en soi une tâche énorme. La mise en œuvre étant une condition nécessaire de l'effectivité du programme, les gestionnaires avaient besoin de telles données pour mieux gérer les programmes et pour répondre à diverses demandes d'information venant du parlement, de cadres de l'administration, d'acteurs locaux, et de professionnel-le-s des médias (Wholey, 1983).

Enfin, certaines préoccupations étaient d'ordre plus intellectuel. Des analystes critiques ont rapidement identifié certains problèmes liés aux programmes sociaux. Ils ont voulu mettre en discussion le processus de résolution des problèmes sociaux qui sous-tendait la conception de ces programmes, critiquer les anciennes hypothèses à cet égard et en développer de nouvelles. Pour cela, ils avaient besoin d'évaluations de la réussite de ces programmes, de données sur les raisons du succès et de l'échec, et ils ont aussi cherché à identifier des modalités par lesquelles les objectifs de ces programmes sociaux auraient pu être atteints par l'intermédiaire d'acteurs non gouvernementaux. Ces objectifs dépassaient l'évaluation de programmes spécifiques, mais une telle évaluation était un prérequis de la démarche.

Qui allait répondre à ces demandes?

Le secteur public manquait de procédures assez développées pour répondre à ces attentes d'évaluation. Il n'y avait pas, dans les programmes sociaux, l'équivalent des fonctions et des professions qui fournissent habituellement ce type de connaissances et de retour sur l'activité dans le secteur privé. Des comptables indépendant-e-s travaillant pour des organisations certifiées vérifient de façon usuelle les comptes des entreprises pour déterminer leur profit et préciser leurs obligations fiscales. Le résultat de leur travail est rendu public, afin que les actionnaires et les agent-e-s de l'administration fiscale puissent juger de la façon dont les entreprises remplissent leurs obligations. Il leur arrive aussi de fournir ce type de données pour des services particuliers d'une entreprise, pour aider à diagnostiquer des problèmes et des réussites, afin d'améliorer le fonctionnement. Des entreprises de conseil en management s'efforcent aussi d'améliorer le fonctionnement des organisations; leur travail va au-delà de la comptabilité et de l'audit financier. Enfin, de nombreuses entreprises ont leurs propres départements de recherche et développement, avec des fonctions de recherche, de développement des produits, et d'essais avant commercialisation. Dans le secteur privé, ces trois fonctions – dresser le bilan des réussites, améliorer le fonctionnement, et développer de nouveaux produits – sont généralement reconnues comme des activités importantes qui dépendent d'un travail d'évaluation. Il n'y a donc rien d'étonnant à ce que des comptables, des consultant-e-s en management, et des spécialistes de recherche et développement, se soient spécialisé-e-s dans ce type d'activités.

Dans le secteur public, on trouve quelques spécialistes qui remplissent ces trois fonctions : des contrôleurs dans l'administration fiscale et budgétaire, des économistes dans les ministères, des analystes de la planification et des systèmes au département de la défense, et des spécialistes du budget partout. Mais ces spécialistes se sont trouvés

dépassés quand il s'est agi d'évaluer les programmes sociaux. L'administration manquait de personnel pour répondre à la demande d'évaluation [...].

De plus, les compétences disponibles au sein de l'administration n'étaient pas toujours pertinentes pour une évaluation des programmes sociaux. Les spécialistes de la planification et des systèmes pouvaient prévoir les effets de nouvelles initiatives, mais avaient peu de formation pour fournir des évaluations rétrospectives du fonctionnement et des conséquences de programmes sociaux existants, à partir des expériences de terrain (Wholey *et al.*, 1970). Des économistes pouvaient facilement fournir une analyse coûts-bénéfices d'un projet hydraulique, mais se sont retrouvé-e-s plus désemparés lorsqu'il s'est agi de mesurer des conséquences d'ordre social telles que l'amélioration de la stabilité familiale ou de l'éducation. Dès lors, les méthodologies existantes en comptabilité, en audit, en enquête quantitative et en prévision menaient à des conclusions plus fragiles lorsqu'on les appliquait au secteur social. De plus, si l'on dispose de théories solides en sciences physiques pour aider à la résolution des problèmes dans les programmes d'ingénierie, et de théories solides en biologie pour aider la médecine, on a beaucoup moins de théories solides dans les sciences sociales. Ceci rend plus difficile la conception de programmes sociaux pour résoudre des problèmes sociaux.

À la fin des années 1960, la demande d'évaluation des programmes sociaux excédait l'offre de personnel ayant des compétences adaptées. Cette demande a attiré vers l'évaluation de nombreux diplômés d'écoles professionnelles et de départements de sciences sociales. L'enseignement supérieur de niveau Master s'est très rapidement développé pendant cette période [...] et le développement des emplois dans le monde académique n'a pas été aussi rapide. [...] Dans ce contexte, la pratique professionnelle de l'évaluation est devenue pour les diplômé-e-s une alternative viable à l'emploi académique. Dès lors, l'évaluation répondait à la fois à un besoin de l'époque, et à l'existence d'une offre de travail, ce qui a conduit au développement d'une profession de l'évaluation.

Les fondements structurels de la profession

L'évaluation est une profession au sens où elle partage certains attributs avec d'autres professions et se distingue en ceci de disciplines purement académiques telles que la psychologie ou la sociologie. Bien qu'elles puissent avoir des fondements et des membres dans le monde académique, les professions se distinguent par le fait qu'elles sont économiquement et socialement structurées de façon à se dédier principalement à l'application pratique des savoirs dans un domaine circonscrit, en mobilisant des financements socialement perçus comme légitimes (Austin, 1981). Les professionnel-le-s ont des contraintes plus fortes que les académiques concernant les activités qu'ils peuvent assumer, et ils développent généralement des normes de pratiques, des codes éthiques, et d'autres contraintes professionnelles. L'évaluation de programmes n'est pas complètement professionnalisée comme le sont la médecine ou le droit; elle ne dispose pas d'un système de licence juridiquement réglementé, par exemple. Mais elle tend vers la professionnalisation plus que d'autres disciplines. Le principal moteur de l'établissement de l'évaluation en tant que pratique professionnelle a probablement été la législation fédérale qui a rendu les évaluations obligatoires et les a financées. [...]

[Si la demande d'évaluation a été forte dans les années 1960 et 1970,] les années 1980 ont vu un déclin dans les financements dédiés à l'évaluation. Selon Cronbach, ce déclin avait commencé à la fin des années 1970 et s'est accentué dans les années 1980, dans le contexte des restrictions budgétaires de l'administration Reagan. Les agences fédérales d'évaluation ont été particulièrement touchées. [...] Mais malgré ce déclin, l'évaluation reste une pratique importante.

En effet, les financements et les activités décrites ci-dessus ont donné à l'évaluation une crédibilité. Sur les effets de l'obligation d'une évaluation au niveau local en éducation, Wholey et White (1973) commentent : « L'impact majeur [de cette obligation légale d'évaluation au niveau local]

a été une acceptation accrue de la démarche d'évaluation; « l'évaluation » est maintenant un terme familier pour les acteurs de l'éducation » (p. 73). Mais les initiatives fédérales n'ont pas seulement apporté à l'évaluation de la crédibilité. De la même façon que la psychologie clinique est devenue une profession par le biais des initiatives fédérales ayant suivi la Seconde Guerre mondiale (Sarason, 1981), les évaluateurs et évaluatrices ont reçu grâce au soutien fédéral l'appui institutionnel nécessaire pour le développement d'une nouvelle profession responsable de certaines tâches (l'évaluation des programmes sociaux) dans un secteur donné (les programmes sociaux publics), avec une base de financement (le financement des fonctions évaluatives) et une légitimité sociale (par le biais du besoin d'évaluation formulé par le gouvernement) (Austin, 1981). Sans ce soutien, l'évaluation aurait pu n'exister que comme une petite discipline académique appliquée, à l'instar de la psychologie communautaire (Heller et Monahan, 1977) ou de la sociologie clinique (Glassner, 1981).

La réponse de l'évaluation

La professionnalisation de l'évaluation. Toute profession se fonde sur un corpus de savoirs unique et transmissible. Le corpus de savoir initial mobilisé par l'évaluation prenait beaucoup appui sur des méthodes et théories préexistantes venant de pratiquement toutes les sciences sociales. Des corpus de savoir plus spécialisés ont émergé au fil de l'approfondissement de l'expérience en évaluation. Cet ouvrage s'intéresse à ces corpus plus spécifiques à l'évaluation. Pour développer et transmettre ces savoirs, certaines universités ont lancé des centres ou des programmes diplômants en évaluation (Evaluation research society, 1980). Selon Wortman (1983a), le premier de ces programmes a ouvert en 1973 – il fait probablement référence au programme doctoral et postdoctoral de formation en évaluation de l'université de Northwestern, où il enseignait. D'autres universités ont formé des évaluateurs par le biais

de formations doctorales adjacentes. Par exemple, entre 1966 et 1986 le Centre pour la recherche en pédagogie et en évaluation des programmes d'enseignement à l'Université de l'Illinois (CIRCE) a formé 49 docteur-e-s spécialisé-e-s en mesure et en évaluation. De nombreux évaluateurs et évaluatrices professionnel-le-s n'ont pas de doctorat, mais la plupart ont au moins un diplôme de Master dans un domaine connexe, comme l'administration publique par exemple (Shadish et Epstein, 1987).

Un autre indice de professionnalisation est la création d'annuaires professionnels, de sociétés savantes, de revues, et de normes professionnelles (Austin, 1981). L'annuaire professionnel du champ était le *Evaluation studies review annual*, publié pour la première fois en 1976 (Glass, 1976) puis annuellement jusqu'en 1987 (Shadish et Reichardt, 1987). Les revues en évaluation existent pour l'ensemble du champ (*Evaluation review*, *Evaluation practice*, *Evaluation and program planning*), mais aussi pour des secteurs spécifiques tels que la santé ou l'éducation (*Evaluation and the health professions*, *Evaluation and educational policy*). En 1976 deux sociétés professionnelles ont été fondées, la société de recherche en évaluation et le réseau en évaluation; en 1985, ils ont fusionné pour former l'Association américaine d'évaluation, avec environ 3 000 membres et des conférences annuelles auxquelles participent entre 500 et 1 000 personnes (American Evaluation Association, 1986). Finalement, des normes de pratique ont été développées, impliquant des seuils de compétences minimales pour les évaluateurs (Rossi, 1982b).

Diversité des pratiques professionnelles. Les évaluateurs et évaluatrices ont répondu aux demandes d'évaluation du gouvernement de trois manières (Cook et Buccino, 1979). Premièrement, certaines entreprises spécialisées dans la recherche contractuelle se sont rapidement spécialisées dans la réponse à des appels d'offres en évaluation. Certaines d'entre elles se sont développées jusqu'à inclure 800 professionnels dotés de doctorat. En 1970, plus de 300 entreprises étaient qualifiées pour recevoir les appels d'offres fédéraux en évaluation (Wholey *et al.*, 1970). Deuxièmement, des chercheurs et des chercheuses au sein des universités ont obtenu des contrats et des financements pour réaliser des

évaluations; certain-e-s ont aussi travaillé comme consultant-e-s auprès de cabinets privés en évaluation. Ces chercheuses et chercheurs ont participé au développement des théories et des méthodes en évaluation. Troisièmement, des services spécialisés en évaluation ont été mis en place au sein des administrations fédérales, étatiques et locales, dans le but de répondre plus rapidement et précisément aux besoins d'information des gestionnaires de programmes.

Dans ces différents cadres, un nombre considérable d'études a été mené en évaluation. […] Ces évaluations étaient très diverses. Elles répondaient à des demandes à différents niveaux de gouvernement. […] Les évaluateurs et évaluatrices ont aussi développé des spécialisations thématiques différentes, incluant l'éducation, la santé publique, la justice pénale, la médecine, la participation au marché du travail, la garantie du revenu, la nutrition, la sécurité routière, l'aide internationale, la santé mentale, et bien d'autres secteurs. Une tâche importante pour ces professionnel-le-s consiste à rester à jour dans leurs domaines de spécialité (Light, 1983; Shadish, et Reichardt 1987).

Mais surtout, les activités menées étaient si différentes qu'il était souvent difficile d'identifier ce qu'elles avaient de commun. Certain-e-s évaluateurs et évaluatrices ont développé des systèmes d'information et de gestion visant à fournir rapidement des données sur le fonctionnement des programmes […]. D'autres ont mené des études de cas et des observations participantes dans la tradition de la sociologie et de l'anthropologie (Guba et Lincoln, 1981). D'autres encore ont lancé des expérimentations sociales de grande échelle, au sein desquelles certaines unités étaient aléatoirement assignées à différents types de traitements, comme dans le cas de l'impôt négatif sur le revenu expérimenté dans le New Jersey (Rossi et Lyall, 1978) […]. Cette diversité a pu conduire un observateur à commenter : « l'évaluation – plus que tout autre science – est-ce que les gens disent que c'est; et les gens actuellement disent que c'est beaucoup de choses différentes » (Glass et Ellett 1980 : 211).

Cette diversité a aussi été alimentée par les financeurs, qui exigeaient des activités de types différents sous la même rubrique d'évaluation. Les évaluateurs ont répondu avec différents cadres et méthodes disciplinaires, et les programmes étudiés avaient par ailleurs des liens théoriques substantiels avec de nombreux champs différents des sciences sociales et divers champs professionnels. Cette diversité persiste à ce jour, faisant en sorte qu'il est difficile pour les évaluateurs et évaluatrices de se mettre d'accord sur ce que devrait être la pratique d'évaluation, et pourquoi. [...] Progressivement, des théories englobantes relatives à l'évaluation des programmes sociaux se sont efforcées d'intégrer cette diversité dans un tout cohérent pour aider les praticiens à comprendre le champ et à améliorer leur pratique.

Les théories en évaluation : de la diversité à l'intégration

[...] L'évaluation est peut-être la spécialité méthodologique la plus large. Sa théorie inclut un large éventail de décisions relatives à la forme, à la conduite et aux effets d'une évaluation. La théorie en évaluation a trait aux méthodes, mais ne s'y limite pas. Pour informer les évaluateurs dans leur choix des méthodes, cette théorie mobilise aussi la philosophie des sciences, l'analyse des politiques publiques, ainsi que des théories relatives à la valeur et à l'utilisation.

Sans ces théories spécifiques, l'évaluation de programmes ne serait qu'un vague ensemble de chercheuses et de chercheurs affilié-e-s à différentes disciplines, et cherchant à appliquer des méthodes de sciences sociales à l'étude des programmes sociaux. Or l'évaluation de programmes est plus que cela, plus que de la méthodologie appliquée. Les évaluateurs et évaluatrices sont progressivement en train de développer un corpus unique de connaissances qui différencient l'évaluation d'autres domaines tout en confirmant sa place à côté de ceux-ci. L'évaluation est diverse

de nombreux points de vue, mais elle trouve un potentiel d'unité intellectuelle dans ce que Scriven appelle « la logique de l'évaluation », qui permet de surmonter les frontières disciplinaires séparant les évaluateurs.

[…] La première tentative d'intégration théorique au sein de ce champ a été le fait de Suchman (1967), dont les idées coïncidaient partiellement avec celles, plus influentes, de Campbell (1969, 1971). Tous deux étaient plus intéressés par le fait de répertorier, à destination des concepteurs des politiques publiques, les réussites des programmes existants dans le secteur public, que par la collecte d'informations pour aider des praticiens au niveau local. Ce qui les intéressait le plus était d'évaluer la pertinence de nouvelles idées qui pourraient être incorporées dans les programmes existants ou dans de nouveaux programmes. La réforme était le mot d'ordre, avec l'idée de tester de nouvelles approches audacieuses; l'évaluation de changements marginaux apportés à des programmes existants était moins valorisée; l'évaluation de pratiques locales pour des raisons locales était globalement négligée.

Avec le temps, les théories en évaluation ont évolué et se sont diversifiées pour refléter l'expérience pratique accumulée. La focale s'est déplacée de l'étude exclusive des résultats à des questionnements plus englobants intégrant la question de la qualité de la mise en œuvre ainsi que les processus causaux médiant l'impact des programmes (Sechrest *et al.*, 1979). Le recours exclusif aux études quantitatives a cédé le pas à des approches plus diversifiées incluant les méthodes qualitatives (Guba et Lincoln, 1981). Alors que les responsables politiques étaient initialement les seuls à l'origine des questions évaluatives et destinataires des résultats, de plus en plus de parties prenantes ont été intégrées à la réflexion (les parties prenantes étant celle pour qui le programme ou son évaluation constitue un enjeu) (Carol H. Weiss, 1983a; 1983b). Aux préoccupations méthodologiques se sont ajoutées des questions relatives au contexte des pratiques d'évaluation, et à la façon de faire une place aux résultats de l'évaluation dans des systèmes très politisés et décentralisés (Cronbach *et al.*, 1980). Les théories actuelles en évaluation couvrent plus

de sujets, ont un meilleur sens des complexités qui pèsent sur la pratique de l'évaluation, et intègrent mieux la diversité de concepts, de méthodes et de pratiques qui traversent le champ (voir par exemple Cronbach, 1982b; Rossi et Freeman, 1985).

[…] Finalement, pourquoi écrire sur les théories en évaluation? Nous le faisons parce qu'il existe un déséquilibre en évaluation entre l'attention importante prêtée aux méthodes et le manque d'attention accordée aux enjeux théoriques qui guident le choix des méthodes. Aucune méthode n'est appropriée partout et toujours. N'utiliser qu'une seule méthode, comme l'expérimentation ou les études de cas, induit des problèmes : par exemple, produire des données moins utiles, ou aboutir à des conclusions inexactes. La théorie en évaluation nous indique quand, où, et pourquoi certaines méthodes devraient être utilisées et d'autres non, en suggérant des séquences au sein desquelles certaines méthodes devraient être appliquées, des manières de combiner différentes méthodes, des types de questions auxquelles une méthode particulière répond plus ou moins bien, et les bénéfices que l'on peut attendre de certaines méthodes par rapport à d'autres. Les théories en évaluation sont comme la stratégie et la tactique militaires, là où les méthodes seraient les munitions et la logistique. Un bon chef ou une bonne cheffe d'armée a besoin de connaître la stratégie et les tactiques pour déployer ses munitions de façon adéquate ou pour organiser la logistique dans diverses situations. Pour la même raison, un bon évaluateur ou une bonne évaluatrice a besoin des théories pour savoir choisir et mettre en œuvre différentes méthodes.

[…] Ceci étant dit, toutes les praticiennes et tous les praticiens de l'évaluation en sont aussi des théoriciennes et théoriciens en herbe. Elles et ils réfléchissent à leur pratique, font preuve de jugement dans le choix de quelles méthodes utiliser dans quelle situation, pèsent les avantages et les inconvénients des choix auxquels elles et ils font face, et apprennent des succès et des échecs de leurs évaluations passées. En pratique, les

concepts pragmatiques développés dans la pratique constituent probablement le fondement le plus solide des théories développées sur le plan académique.

Bibliographie

American Evaluation Association. 1986. « Membership directory (prepared by Evaluation Research Center, School of Education, University of Virginia) ».

Austin, David. 1981. « The development of clinical sociology ». *Journal of Applied Behavioral Science* 17 : 347-50. doi : https://doi.org/10.1177/002188638101700309.

Campbell, Donald T. 1969. « Reforms as experiments ». *American Psychologist* 24(4) : 409-29. doi : https://doi.org/10.1037/h0027982.

Campbell, Donald T. 1971. « Methods for the experimenting society ». *Meeting of the Eastern Psychologial Association*. New York.

Cook, Thomas D., et Alphonse Buccino. 1979. « The social scientist as a provider of consulting services to the federal government ». in *The psychological consultant*, édité par J. J. Platt et R. J. Wicks. New York: Grune & Stratton, p. 103-34.

Cronbach, Lee J., Sueann R. Ambron, Sanford M. Dornbusch, Robert D. Hess, Robert C. Hornik, Denis Charles Phillips, Decker F. Walker et Stephen S. Weiner. 1980. *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass.

Cronbach, Lee Joseph. 1982b. « In praise of uncertainty ». in *Standards for evaluation practice*, édité par P. H. Rossi. San Francisco: Jossey-Bass, p. 49-58.

Cumming, J. H. 1976. « What Congress really wants: A guide to evaluating program effectiveness with examples from the Canadian context ». in *Trends in mental health evaluation*, édité par E. W. Markson et D. F. Allen. Lexington: Rowman & Littlefield, p. 61-70.

Evaluation research society. 1980. *Directory of evaluation training*. Washington, DC: Pintail Press.

Feeley, Malcom, et Austin D. Sarat. 1980. *The policy dilemma: Federal crime policy and the Law Enforcement Assistance Administration, 1968-1978*. Minneapolis: University of Minnesota Press.

Glass, Gene V. 1976. *Evaluation studies review annual*. Vol. 1. Beverly Hills: Sage Publications.

Glass, Gene V., et F. S. Ellett. 1980. « Evaluation research ». in *Annual review of psychology*. Vol. 31, édité par M. R. Rosenzweig et L. W. Porter. Palo Alto: Annual Reviews, p. 211-28.

Glassner, Barry. 1981. « Clinical applications of sociology in health care ». *Journal of Applied Behavioral Science* 17 : 330-46. doi : https://doi.org/10.1177/002188638101700308.

Guba, Egon G., et Yvonne S. Lincoln. 1981. *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. San Francisco: Jossey-Bass.

Heller, Kenneth, et Jonathan T. Monahan. 1977. *Psychology and community change*. Homewood: Dorsey.

House, Ernest R. 1980. *Evaluating with validity*. Beverly Hills: Sage Publications.

Lazarsfled, Paul F., et Morris Rosenberg. 1955. *The language of social research*. Glencoe: Free Press.

Lewin, Kurt. 1948. *Resolving Social Conflicts, Selected Papers on Group Dynamics (1935-1946)*. New York: Harper & Brothers.

Light, Richard J. 1983. *Evaluation studies review annual*. Vol. 8. Beverly Hills: Sage Publications.

Rossi, Peter H. 1982b. *Standards for evaluation practice*. San Francisco: Jossey-Bass.

Rossi, Peter H., et Howard E. Freeman. 1985. *Evaluation: A systematic approach*. 3e éd. Beverly Hills: Sage Publications.

Rossi, Peter H., et K. C. Lyall. 1978. « An overview evaluation of the NIT Experiment ». in *Evaluation studies review annual*. Vol. 3, édité par T. D. Cook, M. L. DelRosario, K. M. Hennigan, M. M. Mark et W. M. K. Trochim. Beverly Hills: Sage Publications, p. 412-28.

Sarason, Seymour B. 1981. « An asocial psychology and a misdirected clinical psychology ». *American Psychologist* 36(8) : 827-36. doi : https://doi.org/10.1037/0003-066X.36.8.827.

Sechrest, Lee, S. G. West, Michael Phillips, R. Render et William Yeaton. 1979. « Some neglected problems in evaluation research: Strength and integrity of treatments ». in *Evaluation studies review annual*. Vol. 4, édité par L. Sechrest et Associates. Beverly Hills: Sage Publications, p. 15-35.

Shadish, William R., et Roberta Epstein. 1987. « Patterns of program evaluation practice among members of Evaluation Research Society and Evaluation Network ». *Evaluation Review* 11 : 555-90. doi : https://doi.org/10.1177/0193841X8701100501.

Shadish, William R., et Charles S. Reichardt. 1987. « The intellectual foundations of social program evaluation: The development of evaluation theory ». in *Evaluation studies review annual*. Vol. 12, édité par C. S. Reichardt et W. R. Shadish. Newbury Park: Sage Publications, p. 13-30.

Suchman, Edward A. 1967. *Evaluative research: Principles and practice in public service and social action programs*. New York: Russel Sage Foundation.

Tyler, Ralph. 1935. « Evaluation: A challenge to progressive education ». *Educational Research Bulletin* 14 : 9-16.

Weiss, Carol H. 1983a. « The stakeholder approach to evaluation: Origins and promise ». in *Stakeholder-based evaluation*, édité par A. S. Bryk. San Francisco: Jossey-Bass, p. 3-14.

Weiss, Carol H. 1983b. « Toward the future of stakeholder approaches in evaluation ». in *Stakeholder-based evaluation*, édité par A. S. Bryk. San Francisco: Jossey-Bass, p. 83-96.

Wholey, Joseph S. 1983. *Evaluation and effective public management*. Boston: Little, Brown.

Wholey, Joseph S., John W. Scanlon, H. G. Duffy, James S. Fukumoto et L. M. Vogt. 1970. *Federal evaluation policy: Analyzing the effects of public programs*. Washington, DC: Urban Institute.

Wholey, Joseph S., et B. F. White. 1973. « Evaluation's impact on Title I elementary and secondary education program management ». *Evaluation* 1 : 73-76.

Wortman, Paul M. 1983. « Evaluation at the frontier: Some 'timely' comments for future use. »

2. La recherche évaluative : principes et pratiques applicables aux services publics et aux programmes sociaux

EDWARD A. SUCHMAN

[Traduit de : Suchman, Edward A. 1967. *Evaluative Research. Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation, extraits des chapitres 1, 2, 3 et 5. Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Russel Sage Foundation.]

Extrait du chapitre 1, p. 7-8 :

Notre approche distinguera « l'évaluation » de la « recherche évaluative ». Nous utiliserons la notion d'évaluation pour désigner de façon générale le processus social consistant à émettre des jugements sur la valeur. Ce processus est fondamental dans pratiquement toutes les activités sociales, qu'elles soient le fait d'individus ou d'organisations complexes. Bien que ce processus implique toujours une forme ou une autre de logique ou de rationalité qui le guide, il ne requiert pas de procédures systématiques visant à réunir et présenter des preuves objectives à l'appui du jugement émis. Dès lors, nous conserverons le terme « évaluation » pour désigner cette évaluation dans son acception courante, renvoyant au processus général d'appréciation et d'évaluation de la valeur.

L'expression « recherche évaluative », en revanche, sera utilisée pour désigner la démarche plus spécifique consistant à utiliser les méthodes et techniques de recherche scientifique à des fins d'évaluation. En ce sens, l'adjectif « évaluative » vient spécifier un type de recherche. L'accent est d'abord placé sur le nom « recherche », et la recherche évaluative renvoie

aux procédures de collecte et d'analyse des données qui augmentent les chances de pouvoir « prouver » plutôt que simplement « affirmer » qu'une activité sociale a telle ou telle valeur.

Extrait du chapitre 2, p. 20-23 :

[…] La recherche évaluative représente une tentative d'utiliser des méthodes scientifiques afin de déterminer l'utilité d'une activité. Dans la définition du protocole de recherche comme dans les procédures de collecte et d'analyse des données, la recherche évaluative doit s'efforcer d'adhérer le plus possible aux canons de la méthode scientifique. […] En adoptant une méthode scientifique, on espère que les résultats de l'étude évaluative seront plus objectifs, et qu'on se donnera la possibilité d'en vérifier la fiabilité et la validité.

Il ne fait aucun doute que plus une étude évaluative sera conçue et mise en œuvre en suivant les règles de la méthode scientifique, plus on pourra être confiant quant à l'objectivité des résultats produits. Toutefois, il ne faut pas oublier que la recherche fondamentale (*basic research*) a un objectif différent de celui de la recherche évaluative. […] La recherche fondamentale a pour objectif premier l'établissement d'un savoir, le fait de prouver ou d'invalider une hypothèse. Il n'est généralement pas attendu qu'elle soit suivie d'une action particulière de la part d'une administration. Le critère essentiel de détermination du « succès » d'un projet de recherche fondamentale réside dans la validité scientifique des résultats auxquels il aboutit, validité qui est évaluée sur la base des règles de la méthode scientifique.

La recherche évaluative, en revanche, est généralement une recherche de type appliqué ou administratif (*administrative research*), dont l'objectif premier est de déterminer le degré auquel un programme ou une procédure donné-e aboutit à tel ou tel résultat souhaité. Le « succès » d'un projet d'évaluation dépendra alors largement de son utilité pour favoriser, pour l'administration concernée, une amélioration du service

fourni. Dès lors, même si des critères scientifiques entrent en jeu pour déterminer le degré de confiance que l'on peut accorder aux résultats d'une étude évaluative, l'utilité de cette dernière dépendra surtout de critères d'ordre administratif. Contrairement au chercheur en recherche fondamentale, le chercheur en recherche appliquée doit constamment garder à l'esprit la question de l'utilité potentielle de ses résultats. Il est rare qu'il puisse se satisfaire de l'idée selon laquelle « l'opération a été un succès même si le patient est décédé ».

C'est précisément ce qui rend la recherche évaluative « difficile ». En théorie, la définition du protocole de la recherche pose généralement moins problème que dans la recherche non-évaluative. Les « hypothèses » découlent largement des objectifs et des procédures mis en avant par le programme ou le service que l'on cherche à évaluer (même si, comme nous le verrons, en pratique leur identification ne va souvent pas de soi). De plus, le protocole de recherche essaie pratiquement toujours de se conformer à un modèle de type expérimental, impliquant des mesures avant et après sur un groupe expérimental et un groupe de contrôle (même si, là aussi, des adaptations sont possibles). Concevoir un dispositif « idéal » d'évaluation est probablement un des exercices de recherche les plus simples que l'on puisse imaginer.

Ce ne sont pas tant les principes de la recherche qui rendent les études évaluatives difficiles, que la difficulté pratique à rester fidèle à ces principes face à des considérations administratives. De façon beaucoup plus nette que pour le chercheur ou pour la chercheuse en recherche fondamentale, l'évaluateur ou l'évaluatrice perd le contrôle de la situation de recherche. Les objectifs du programme à évaluer ont déjà été définis par quelqu'un d'autre, le plus souvent par un administrateur ou une administratrice de ce programme qui a un intérêt direct dans celui-ci. Il est à la fois difficile et pénible de le forcer à remettre en question les hypothèses sous-jacentes à ces objectifs. Le fait de diviser ces objectifs en plusieurs étapes intermédiaires en vue d'un objectif de plus long terme paraîtra souvent à l'administrateur ou à l'administratrice du programme comme une tentative de limiter la portée de ce programme, voire de le

détruire. L'introduction d'un critère de performance plutôt que d'effort, ainsi que de critères d'effectivité et d'efficacité, pourront lui paraître comme une remise en cause de sa compétence. La présence de ce tiers biaisé entre l'évaluateur ou l'évaluatrice et son objet de recherche induit des difficultés largement inévitables, que ne rencontre généralement pas le chercheur ou la chercheuse en recherche fondamentale.

À la suspicion naturelle et à l'opposition de l'administrateur ou de l'administratrice du programme, il faut ajouter le fait que la plupart des études évaluatives supposent un certain degré d'interférence avec les activités en cours, ce qui induit des motifs supplémentaires de critiques et de refus de coopération. L'évaluateur ou l'évaluatrice ne se limite généralement pas à l'observation et à la mesure comme le fait le chercheur ou la chercheuse en recherche fondamentale; il ou elle doit souvent demander une modification des procédures voire une interruption complète du service, afin de garantir l'existence d'une forme de groupe de contrôle ou de comparaison. Il ou elle peut solliciter la collecte et l'enregistrement d'informations supplémentaires non nécessaires au fonctionnement du programme mais essentiels pour l'évaluation. Et tout cela, pour faire sens, doit généralement être fait au fil du fonctionnement courant du programme. En effet il fait peu sens, dans le cadre d'une évaluation, de mettre en place une situation expérimentale complètement artificielle.

Du point de vue de l'évaluateur ou de l'évaluatrice, l'évaluation doit être « simple et pratique », alors même que les prérequis méthodologiques de la recherche évaluative se complexifient et se spécialisent. […] Il s'agit pour la recherche évaluative de trouver un équilibre entre les limites administratives et les exigences méthodologiques. Toutes les évaluations ne requièrent pas le même niveau de rigueur scientifique, et beaucoup de décisions administratives se prennent sur la base d'évaluations limitées. Il s'agira, à l'avenir, de mieux comprendre les différentes formes que peut prendre l'évaluation, et quand chacune est adaptée.

Finalement, l'évaluation souffre aujourd'hui d'un manque général de financements, de structures, de temps et de personnel. Les services publics et les associations chargées de l'action sociale[1] s'activent à développer des interventions visant à répondre aux besoins les plus criants des populations locales. La plupart de ces programmes ont une validité apparente et semblent avoir rarement besoin de la recherche évaluative pour prouver leur efficacité. Ce n'est que récemment qu'une plus grande priorité a été donnée à la recherche évaluative [...]. Mais si on les compare aux ressources dédiées aux interventions elles-mêmes, et même à la recherche fondamentale, la quantité de temps, d'argent et de personnel allouée à l'évaluation est tristement inadéquate. Comme toute recherche, la recherche évaluative coûte de l'argent, et pas plus que la recherche fondamentale, on ne peut en faire une activité adjacente ou à temps partiel des personnels chargés des programmes. Même si on peut intégrer dans les services existants quelques collectes d'informations supplémentaires fournissant des données de base pour l'évaluation, cela ne suffit pas en soi à fournir des évaluations de programmes valides. Des personnels de recherche formés, utilisant des fonds d'évaluation spécifiques, sont la condition sine qua non du développement d'une recherche évaluative « scientifique ». Tant que ce prérequis de base n'est pas rempli, on peut douter que la recherche évaluative du futur soit de meilleure qualité que celle du passé.

On peut légitimement se demander si le champ actuel de la recherche évaluative est prêt à occuper une place plus conséquente. On manque d'analyses critiques sur la place souhaitable de l'évaluation, et on manque encore plus de discussions méthodologiques sur les adaptations des protocoles de recherche qui permettraient de rendre les études évaluatives plus productives. Difficile de dire qui est la poule et qui est l'œuf. Est-ce que la recherche évaluative reste sous-investie parce que

1. NdT : Community action agencies, ensemble d'institutions publiques et d'associations locales impliquées dans la mise en œuvre du Community action program, un des volets de la politique de lutte de la pauvreté développée en 1964 par l'administration Johnson.

nous n'en savons pas encore assez sur les façons de faire de bonnes études évaluatives, ou est-ce que nous n'avons pas assez appris justement parce que la recherche évaluative a été tant négligée? Il manque encore à l'évaluation, nous en sommes convaincus, une analyse systématique des principes théoriques, méthodologiques et administratifs qui sous-tendent ses objectifs et ses procédures. Une des raisons essentielles du discrédit qui pèse aujourd'hui sur les études évaluatives est l'absence de principes et de normes clairs pour leur conduite.

Extrait du chapitre 3, p. 31 :

L'évaluation [peut être définie] comme la détermination (qu'elle soit basée sur des opinions, des documents, des données subjectives ou objectives) des résultats (qu'ils soient désirables ou indésirables, éphémères ou durables, immédiats ou tardifs) produits par une activité (qu'il s'agisse d'un programme ou d'une partie d'un programme, d'un médicament, d'un traitement, d'une action récurrente ou ponctuelle) conçue pour réaliser un objectif valorisé (qu'il soit large ou précis, de long, moyen ou court terme, qu'il soit formulé en termes d'effort ou de performance). Cette définition comprend quatre dimensions clés : 1) un processus – la « détermination », 2) un critère – les « résultats », 3) un stimulus – « l'activité », et 4) une valeur – « l'objectif ». La méthode scientifique, avec ses techniques de recherche spécifiques, fournit le moyen le plus prometteur de « déterminer » la relation entre le « stimulus » et « l'objectif », selon des « critères » mesurables.

Cela ne revient pas à exclure les méthodes « non-scientifiques » d'évaluation, même si l'utilisation d'une méthodologie « scientifique » est plus valorisée. L'accent est ici placé sur le processus d'évaluation en général en tant que but, et non sur la recherche évaluative qui ne constitue qu'un des moyens d'atteindre ce but. La conception, le développement et la mise en œuvre des programmes soulève de nombreuses questions de type évaluatif auxquelles on peut répondre sans

avoir recours à la recherche, et beaucoup de questions auxquelles, en l'état de nos connaissances, même les meilleures techniques de recherche ne permettent pas de répondre. La recherche évaluative est un outil, et comme tout outil, pour être pleinement utile, elle doit être conçue pour une fonction spécifique. La dernière recommandation que nous ajouterons est la suivante : l'évaluateur ou l'évaluatrice doit être conscient-e du type d'outil qu'il ou elle utilise, et si l'évaluation requiert une approche de recherche scientifique, il ou elle doit veiller à ne pas lui substituer une appréciation subjective. Nous sommes convaincus de la nécessité actuelle de développer une recherche évaluative plus scientifique, et que le progrès en évaluation passera d'abord par un examen plus approfondi des objectifs de chaque programme en incluant les hypothèses sous-jacentes, ensuite par le développement de critères mesurables correspondant à chacun de ces objectifs, et enfin par la mise en place d'une situation de contrôle pour déterminer le degré auquel ces objectifs sont atteints et identifier les effets secondaires éventuels. Ce sont là les trois conditions *sine qua non* d'une recherche évaluative qui relève vraiment de la recherche et pas simplement du jugement subjectif.

Extrait du chapitre 5, p. 84 :

[…] Les différences entre recherche fondamentale et recherche appliquée ont fait couler beaucoup d'encre; on les représente souvent comme les pôles opposés d'un *continuum* allant de la recherche « pure » à la recherche « d'ingénierie ». Certes, cette controverse est plus polémique que productive, et une grande partie sinon la majorité de la recherche a des éléments à la fois fondamentaux et appliqués. Pour autant, il existe bien une différence essentielle entre les deux démarches, notamment aux deux extrémités du *continuum*. La confusion entre les deux sous-tend l'essentiel des accusations dénonçant la recherche évaluative comme « non valide » ou la recherche fondamentale comme « inutile ».

De façon générale, la recherche évaluative et la recherche non-évaluative se distinguent par l'accent différent qu'elles placent sur les objectifs et les méthodes. En termes d'objectifs, la recherche évaluative est plus susceptible de viser un but pratique – elle est particulièrement préoccupée par son utilité. Cette recherche, si elle atteint son but, devrait permettre de fournir des informations utiles à la conception, au développement et à la mise en œuvre des programmes. Comme l'indique Fleck, « Le trait caractéristique qui transforme une quête de savoir en un projet d'évaluation est la présence d'un objectif selon lequel le savoir produit doit servir de guide à une action pratique » (Fleck, 1963 : 717).

Par contraste, la recherche non-évaluative, bien qu'elle puisse avoir des implications pratiques, vise d'abord une meilleure compréhension, plutôt que la manipulation du réel ou l'action. Un projet de recherche fondamentale a pour objectif essentiel la quête d'un savoir nouveau, indépendamment de la valeur de ce savoir pour produire un changement social. L'accent est mis sur l'étude les relations entre variables plutôt que sur la capacité humaine d'influencer ces relations par des interventions contrôlées.

Un corollaire de cette distinction entre compréhension et manipulation renvoie à des degrés contrastés d'abstraction ou de spécificité. La recherche fondamentale vise la formulation de généralisations théoriques ou de prévisions abstraites, alors que la recherche appliquée met l'accent sur l'action, dans une situation très spécifique impliquant des prévisions concrètes.

Bibliographie

Fleck, Andrew C. 1963. « Evaluation research programs in public health practice ». *Annals of the New York Academy of Sciences* 107. doi : 10.1111/j.1749-6632.1963.tb13314.x.

3. Des différences entre l'évaluation et la recherche, et de leur importance

SANDRA MATHISON

[Traduit de : Mathison, Sandra. 2008. « What Is the Difference between Evaluation and Research-and, Why Do We Care? » in *Fundamental Issues in Evaluation*, edited by N. Smith and P. Brandon. (extraits p.188-195). Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Guilford Publications.]

La question des différences entre l'évaluation et la recherche découle du constat que l'évaluation, en tant que discipline, s'appuie sur d'autres disciplines pour ses fondements, et en particulier sur les sciences sociales pour ses méthodes. Au fil de l'évolution de l'évaluation en tant que discipline et en tant que profession, cette question est parfois posée pour clarifier ce qui la distingue. Cette délimitation d'une profession d'évaluation est également liée à une discussion sur qui évalue et qui peut évaluer. Quelles connaissances et compétences l'évaluation exige-t-elle, et en quoi diffèrent-elles de celles des chercheurs en sciences sociales? […]

Pourquoi se soucier de cette distinction?

La pratique de l'évaluation est ancienne, mais l'évaluation en tant que discipline et en tant que profession est assez récente. En effet, l'évaluation en tant que profession et la clarification concomitante de ses contours en tant que discipline ne datent que de quatre ou cinq décennies. Cette nouveauté induit une certaine hésitation lorsqu'il s'agit de caractériser

ce que nous faisons lorsque nous disons faire de l'évaluation (et non de la recherche), et un effort d'explicitation des principes fondamentaux de l'évaluation (qui la distinguent de la recherche). Les disciplines sont des domaines d'études ayant leur propre logique et qui impliquent diverses théories permettant d'étudier ces domaines. La discipline de l'évaluation se caractérise par une logique particulière (Fournier, 1995; Scriven, 1999) et par des sous-théories, notamment relatives à l'attribution de la valeur, à la pratique, à la prescription et à l'utilisation. (Pour une discussion plus complète de ces sous-théories qui constituent la discipline de l'évaluation, voir (Mathison, 2004b) [...] On pourrait imaginer que des questions telles que « Quelle est la différence entre les statistiques (une discipline plus récente) et les mathématiques (une discipline plus établie)? » ont aussi été posées à l'époque où la discipline des statistiques est peu à peu venue s'ajouter aux usages plus communs des probabilités. La question de la différence entre l'évaluation et la recherche pousse de façon productive les théoricien-ne-s et les praticien-ne-s de l'évaluation à réfléchir et à décrire les fondements de leur discipline, d'abord par comparaison avec la recherche en sciences sociales, mais aussi de plus en plus dans le cadre d'une exploration analytique de l'évaluation en tant que telle. [...]

L'incapacité des méthodes de recherche en sciences sociales à répondre à elles seules aux questions sur la valeur des programmes a entraîné une croissance considérable de l'évaluation en tant que discipline distincte. [...] La question des différences entre l'évaluation et la recherche renvoie aussi aux connaissances et aux compétences dont ont besoin les évaluateurs, et notamment à la question de savoir s'il existe des domaines de connaissances et de compétences qui leur sont spécifiques. La connaissance et la maîtrise des méthodes de recherche en sciences sociales sont utiles, mais elles ne suffisent pas, que ce soit sur le plan du répertoire méthodologique ou du type de connaissances et de compétences nécessaires en matière d'évaluation. Scriven suggère que les évaluatrices et les évaluateurs doivent également savoir comment rechercher les effets non anticipés et adjacents, comment déterminer

la valeur en fonction de différents points de vue, comment traiter les questions et les valeurs controversées et comment faire la synthèse des faits et des valeurs (Coffman, 2003).

Bien qu'il y ait de plus en plus de programmes d'études supérieures pour former les futurs évaluatrices et évaluateurs, il n'en demeure pas moins que de nombreuses personnes en viennent à ce métier par des chemins détournés, souvent avec une connaissance des méthodes et des statistiques des sciences sociales, mais relativement peu de maîtrise des domaines de connaissances décrits par Scriven. Dans ces conditions, il est naturel que cette question continue de se poser, alors que de plus en plus d'évaluatrices et d'évaluateurs novices font leurs premiers pas dans une profession qui exige qu'elles et ils acquièrent des connaissances et des compétences supplémentaires.

Quelles sont les différences entre l'évaluation et la recherche?

Bien que l'on entende parfois que l'évaluation ne diffère pas de la recherche, en particulier de la recherche appliquée en sciences sociales, et bien que les deux démarches soient effectivement liées, les évaluateurs revendiquent une différence. Parce que l'évaluation exige l'étude de ce qui est, elle exige de faire de la recherche. La détermination de la valeur ou du mérite d'un sujet évalué nécessite une certaine connaissance factuelle dudit sujet, voire même de sujets similaires. Toutefois, l'évaluation exige plus que des faits sur le sujet évalué. [...] L'évaluation exige également la synthèse des faits et des valeurs dans la détermination du mérite et de la valeur. En revanche, la recherche examine les connaissances factuelles, mais ne fait pas nécessairement intervenir les valeurs et n'inclut donc pas nécessairement un processus d'évaluation.

Même si la recherche et l'évaluation sont liées entre elles, on tente souvent de trouver une manière de les distinguer. Par exemple, certain-ne-s considèrent la recherche comme un sous-ensemble de l'évaluation; d'autres considèrent l'évaluation comme un sous-ensemble de la recherche; d'autres encore considèrent l'évaluation et la recherche comme les deux extrêmes d'un *continuum*; et certains, enfin, considèrent l'évaluation et la recherche comme un diagramme de Venn avec un recoupement partiel entre les deux. Évaluation et recherche sont le plus souvent comparées du point de vue des objectifs de chacune (c'est-à-dire le résultat anticipé de la recherche ou de l'évaluation), des méthodes d'enquête utilisées et des critères selon lesquels on juge de leur qualité.

L'objet de l'évaluation et de la recherche

La liste ci-dessous illustre la pléthore des critères souvent utilisés pour distinguer l'évaluation de la recherche.

- L'évaluation s'intéresse au cas particulier; la recherche vise la généralisation.
- L'évaluation vise à améliorer les choses, tandis que la recherche a pour but de prouver quelque chose.
- L'évaluation fournit un appui à la prise de décision; la recherche sert de base pour tirer des conclusions.
- Évaluation – et maintenant? La recherche… Qu'en est-il?
- Évaluation – à quel point cela fonctionne? La recherche – comment cela fonctionne?
- L'évaluation porte sur la valeur; la recherche, sur ce qui est.

Ces tentatives de distinction directe entre l'évaluation et la recherche sont problématiques, car elles caricaturent les deux afin de mettre en évidence des différences claires. […]

Prenons, par exemple, la distinction fréquente entre généralisation et particularisation – ni l'une ni l'autre démarche ne suffisant à résumer la recherche ni l'évaluation. Bien que l'évaluation s'intéresse foncièrement au cas particulier en ce sens qu'elle se concentre sur un sujet évalué, ses résultats peuvent néanmoins donner lieu à des montées en généralité, et le font souvent. Le traité de Cronbach (1982) sur la conception des évaluations traitait spécifiquement de la question de la validité externe ou de la généralisation des résultats. Le type de généralisation décrit par Cronbach n'était pas une affirmation sur une population fondée sur un échantillon, mais plutôt une revendication de connaissance fondée sur le constat de similitudes entre UTOM (unités, traitements, observations, milieux). Cela suggère qu'une personne externe à cette évaluation, mais dans un contexte similaire à celui du sujet évalué, pourrait tirer la même conclusion que l'évaluateur. [...]

Réciproquement, la recherche ne vise pas nécessairement ni principalement la généralisation. Une analyse historique des causes de la Révolution française, une ethnographie des Minangkabaus, ou une étude écologique des îles Galápagos ne sont pas forcément réalisées dans le but de généraliser les résultats à toutes les révolutions, à toutes les cultures matriarcales ou à tous les écosystèmes autonomes. [...]

Une autre distinction fréquente consiste à affirmer que l'évaluation vise la prise de décision, tandis que la recherche vise à établir ou confirmer une conclusion. Mais l'évaluation ne vise pas nécessairement la prise de décision ou l'action. Elle peut aussi être un but en soi, sans être conçue comme le préalable à une décision, à un changement ou à une amélioration. [...] La discipline de l'évaluation distingue clairement les conclusions évaluatives des recommandations. Bien que logiquement liées, ces deux activités constituent en fait des formes de raisonnement distinctes.

Inversement, certaines formes de recherche en sciences sociales sont étroitement liées à l'action ou à la recherche de solutions aux problèmes sociaux. Par exemple, diverses formes de recherche-action visent à

atténuer les problèmes, à prendre des mesures, à déterminer ce qui est valorisé et ce qui ne l'est pas, et à œuvrer en faveur d'une réalité plus conforme à ces valeurs. […]

Certes, le lien entre la recherche en sciences sociales et l'action, y compris la prise de décision, peut être moins étroit que pour l'évaluation. Une grande partie de la recherche en sciences sociales aspire à influencer les politiques et les pratiques, mais le fait souvent de manière indirecte et par l'accumulation de connaissances issues de la recherche – ce qui pourrait être considéré comme une influence au niveau macro. En pratique, les résultats d'une étude singulière sont généralement considérés comme insuffisants pour servir directement de base à une décision ou un plan d'action : ils fournissent un tableau incomplet. Toutefois, les spécialistes des sciences sociales espèrent que les études qu'ils mènent influeront sur la prise de décision en sensibilisant les décideurs à un enjeu, en contribuant à un ensemble de recherches qui, dans leur cumulativité, contribuent à éclairer la prise de décision, identifier des alternatives politiques, informer les décideurs et décideuses politiques, et permettent aux interventions de prendre appui sur des résultats issus de la recherche (d'où l'intérêt pour les pratiques fondées sur des données probantes). L'évaluation peut avoir une incidence plus directe sur les décisions concernant des sujets évalués spécifiques, c'est-à-dire influencer les décisions au niveau micro. Cependant, la préoccupation de la discipline, depuis des décennies, à l'égard de l'utilisation de l'évaluation (ou son absence), suggère que le lien entre conclusions évaluatives et prise de décision ne va pas non plus de soi. […]

Bien que l'évaluation soit plus susceptible de contribuer à la prise de décision au niveau micro et la recherche d'informer la décision à un niveau plus macro, cette distinction résiste mal à un examen plus attentif.

Différences dans les méthodes de recherche et d'évaluation en sciences sociales

L'évaluation, surtout à ses débuts en tant que discipline, a largement emprunté ses méthodes d'enquête aux sciences sociales. Les premier-e-s évaluateurs et évaluatrices ont été formé-e-s dans les traditions des sciences sociales – en particulier la psychologie et la sociologie, et dans une moindre mesure l'anthropologie – et s'en sont donc inspiré-e-s dans leurs pratiques d'établissement des preuves empiriques. Pour certain-e-s, cela n'a pas changé, mais pour beaucoup d'autres, la pratique de l'évaluation a évolué de façon importante. Étant donné que l'évaluation traite nécessairement de questions telles que les besoins, les coûts, l'éthique, la faisabilité et légitimité, les évaluateurs utilisent un éventail de stratégies d'établissement de la preuve beaucoup plus large que les sciences sociales. En plus de tous les moyens mobilisés par les sciences sociales pour établir des connaissances, les évaluatrices et les évaluateurs peuvent emprunter à d'autres disciplines telles que la jurisprudence, le journalisme, les arts, la philosophie, la comptabilité et l'éthique.

La crise épistémologique qui a élargi le répertoire des stratégies acceptables pour la collecte et l'analyse des données en sciences sociales ne s'est pas jouée de la même façon en évaluation. Bien que la profession de l'évaluation ait exploré le débat entre méthodes quantitatives et qualitatives, et soit attentive à l'hégémonie du modèle des essais cliniques avec assignation aléatoire, l'évaluation, en tant que pratique, a librement emprunté à toutes les disciplines et à tous les modes de pensées pour travailler à la fois les faits et les valeurs. J'utilise dans une autre publication la notion de l'épistémologie anarchiste de Feyerabend pour décrire cette tendance de l'évaluation (Mathison, 2004a). L'anarchisme est le rejet de toutes les formes de domination. Ainsi, utiliser une épistémologie anarchiste en évaluation implique de rejeter toute domination d'une méthode sur les autres, d'une idéologie unique, d'une idée unique de progrès; c'est un refus du chauvinisme scientifique; de la bien-pensance

des intellectuels; de la prévalence des évaluateurs sur les ressortissants et les fournisseurs de services; de la supériorité du texte académique sur les traditions orales et les autres traditions écrites; c'est se méfier des certitudes.

La pratique de l'évaluation n'implique pas d'adopter à la lettre des modes de connaissance spécifiques issus des sciences sociales, mais plutôt de réfléchir à la façon la plus adéquate d'évaluer en fonction du contexte. Les exemples en sont nombreux, mais deux illustrations suffiront : la « technique du changement le plus significatif » et le « PhotoVoice ». Bien que ces deux méthodes d'enquête puissent être utilisées en sciences sociales, elles revêtent une importance particulière en évaluation par leur valorisation du point de vue des parties prenantes (une perspective unique à l'évaluation et non partagée par la recherche en sciences sociales).

La technique du changement le plus significatif « implique la collecte d'histoires de changement significatif (CS) provenant du terrain, et la sélection systématique des plus significatives de ces histoires par des panels composés de parties prenantes ou d'intervenant-e-s. Ces panels ont d'abord pour mission de 'rechercher' l'impact du projet. Une fois les changements identifiés, diverses personnes se réunissent, lisent les histoires à voix haute et mènent des discussions régulières et souvent approfondies sur la valeur des changements signalés. Lorsque la technique est mise en œuvre avec succès, des équipes entières se focalisent sur l'impact du programme » (Davies et Dart, 2005).

Le deuxième exemple est PhotoVoice, une stratégie communautaire visant à impliquer les parties prenantes dans le changement social, en les amenant à identifier ce qu'elles valorisent et ce qu'elles estiment (Wang, Yuan et Feng, 1996). PhotoVoice utilise des techniques de photographie documentaire pour permettre aux « bénéficiaires de services » ou aux « sujets » de prendre le contrôle de leur propre représentation; elle a été utilisée avec des réfugié-e-s, des immigrant-e-s, des sans-abris et des personnes handicapées. PhotoVoice vise à mettre à profit les

connaissances et les valeurs personnelles et à favoriser le renforcement de la capacité d'évaluation; les individus acquièrent une compétence qui leur permet de continuer à être une voix au sein de leur collectivité. [...]

Les approches en évaluation sont souvent liées à des traditions particulières des sciences sociales, de sorte que l'on néglige parfois l'éventail plus large des méthodes d'enquête mobilisées, qui font la spécificité de l'évaluation. Les deux exemples ci-dessus (technique de changement le plus significatif et PhotoVoice) illustrent comment les évaluatrices et les évaluateurs ont commencé à élaborer des méthodes d'évaluation spécifiques pour juger de manière adéquate et appropriée le mérite ou la valeur des sujets évalués. Il existe de nombreux autres exemples de ce type : l'évaluation coordonnée[1] (*cluster evaluation*), l'évaluation rapide en milieu rural (*rapid rural appraisal*), l'étude d'évaluabilité (*evaluability assessment*) et la méthode centrée sur les cas de succès (*success case method*) – pour n'en citer que quelques-uns.

Les critères pour juger de la qualité de l'évaluation et de la recherche

Une autre manière d'envisager la différence entre l'évaluation et la recherche – soulignée notamment par Michael Quinn Patton dans la discussion EVALTALK[2] de 1998 sur cette question – part des critères utilisés pour juger de la qualité des travaux. Pour Patton, les critères différents mobilisés en recherche et en évaluation résultent des objectifs différents des deux démarches. L'objectif premier de la recherche est de contribuer à la compréhension du fonctionnement du monde, de sorte

1. NdT : Cette démarche consiste à organiser un échange entre les évaluateurs individuels d'un ensemble de projets, afin d'identifier des problématiques communes.
2. NdT : Liste de discussion en évaluation.

que la recherche est jugée à l'aune de son exactitude, en fonction de sa validité perçue, de sa fiabilité, de l'attention portée à la causalité et de son caractère généralisable. L'évaluation est également jugée par son exactitude, mais aussi par son utilité, sa faisabilité et sa pertinence. […]

Une caractéristique importante de l'évaluation est la place centrale accordée au point de vue des parties prenantes, sans équivalent dans la recherche en sciences sociales. Les évaluations sont jugées en fonction de si et comment les points de vue des parties prenantes sont pris en compte. Bien que basés sur différents fondements épistémologiques, tous les modèles utilisés en évaluation intègrent d'une manière ou d'une autre le point de vue des parties prenantes – les évaluations, les valeurs et les significations véhiculées par les parties prenantes sont des éléments essentiels de toute évaluation. La recherche en sciences sociales peut intégrer la perspective des parties prenantes, mais c'est loin d'être systématique. Lorsque les participantes et participants à la recherche sont désignés comme des parties prenantes, c'est souvent pour faire référence aux personnes auprès de qui les données sont collectées, plutôt que pour prendre véritablement en considération leurs intérêts. L'extrait suivant tiré d'un document publié par les Centres de contrôle des maladies (*Centers for disease control*) aux États-Unis, le document-cadre pour l'évaluation de programme (*Framework for Program Evaluation*) (1999), illustre le caractère central de cette démarche d'inclusion des parties prenantes dans l'évaluation. Aucune approche dans la recherche en sciences sociales n'inclut ce concept de manière aussi fondamentale.

> Le cycle d'évaluation commence par l'implication des parties prenantes (i.e. les personnes ou les organismes qui ont un intérêt dans ce qui sera appris et dans ce qui sera fait d'une évaluation). Les travaux de santé publique impliquent des partenariats; par conséquent, toute évaluation d'un programme de santé publique exige de tenir compte des systèmes de valeurs des partenaires.

Les parties prenantes doivent participer à l'enquête pour s'assurer que leurs points de vue sont compris. Lorsque les parties prenantes ne sont pas impliquées, les conclusions de l'évaluation risquent d'être ignorées, critiquées ou rejetées, car elles ne répondent pas aux questions que se posent les parties prenantes ou sont éloignées de leurs valeurs. Après avoir été impliquées dans l'enquête, les parties prenantes peuvent aider à exécuter les autres étapes. Il est essentiel d'identifier et de faire participer les trois groupes suivants :

1. Les personnes impliquées dans le fonctionnement du programme (par exemple les financeurs/-ceuses, les collaborateurs/-trices, les partenaires, les administrateurs/-trices, les gestionnaires et le personnel)

2. Les individus desservis ou touchés par le programme (par exemple les ressortissantes et ressortissants, leurs familles, des organisations locales, des établissements universitaires, des élu-e-s, des collectifs militants, des associations professionnelles, des sceptiques, des opposant-e-s, le personnel d'organisations connexes ou concurrentes…)

3. Les principaux utilisateurs et utilisatrices de l'évaluation (par exemple, les personnes qui sont en mesure de prendre des décisions concernant le programme). Dans la pratique, les utilisatrices et utilisateurs primaires constituent un sous-ensemble des parties prenantes. Une évaluation réussie les identifiera dès le début de son élaboration et maintiendra des interactions fréquentes avec elles et eux afin que l'évaluation tienne compte de leurs valeurs et réponde à leurs besoins d'information spécifiques.

Conclusion

L'évaluation et la recherche diffèrent – une différence de degré sur les *continuums* particularisation-généralisation et démarche orientée vers la décision ou vers l'analyse, pour reprendre les critères les plus souvent retenus. Mais elles diffèrent aussi sur le plan des méthodes : l'évaluation inclut les méthodes de collecte et d'analyse des données issues des sciences sociales, mais, en tant que discipline, elle a aussi développé des méthodes spécifiques. Le jugement sur la qualité des travaux mobilise par ailleurs des critères différents en recherche et en évaluation : l'exactitude compte dans les deux cas, mais l'évaluation mobilise aussi des critères spécifiques relatifs à l'utilité, à la faisabilité, à la pertinence et à l'inclusion des parties prenantes.

À mesure que l'évaluation progresse en tant que discipline, avec une vision plus claire de son objectif spécifique, la question de sa distinction avec la recherche pourrait s'estomper. Cette question restera toutefois fondamentale tant que la méthodologie de l'évaluation continuera à recouper considérablement les méthodes des sciences sociales et tant que les évaluatrices et évaluateurs viendront à cette profession à partir de formations classiques de sciences sociales. Comme nous l'avons suggéré plus haut, ces questions fondamentales sont l'occasion de clarifier ce qu'est l'évaluation en tant que pratique, profession et discipline.

Bibliographie

Centers for Disease Control. 1999. « Framework for program evaluation in public health ».

Coffman, Julia. 2003. « Michael Scriven on the differences between evaluation and social science research ». *The Evaluation Exchange* 9(4).

Cronbach, Lee J. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Davies, Rick, et Jess Dart. 2005. « The most significant change technique: A guide to its use » https://mande.co.uk/docs/MSCGuide.pdf.

Fournier, Deborah. 1995. *Reasoning in evaluation: Inferential links and leaps*. San Francisco: Jossey-Bass.

Mathison, Sandra. 2004a. « An anarchist epistemology in evaluation ». *Annual meeting of the American Evaluation Association*. Atlanta.

Mathison, Sandra. 2004b. « Evaluation theory ». in *Encyclopedia of evaluation*, édité par S. Mathison. Newbury Park: Sage Publications, p. 142-43.

Scriven, Michael. 1999. « The nature of evaluation: Part I. Relation to psychology ». *Practical Assessment, Research and Evaluation* 6(11). doi : https://doi.org/10.7275/egax-6010.

Wang, Caroline, Yan L. Yuan et Ming L. Feng. 1996. « Photovoice as a tool for participatory evaluation: The community's view of process and impact ». *Journal of Contemporary Health* 4 : 47-49.

4. La malédiction de l'évaluation au sein des universités

GARY B. COX

[Traduit de : Cox, Gary B. 1990. « La malédiction de l'évaluation au sein des universités » « On the Demise of Academic Evaluation. » *Evaluation and Program Planning* 13(4) : 415–19 (extraits). Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation d'Elsevier.]

Les départements universitaires n'ont pas historiquement apporté un grand soutien à l'évaluation de programmes. Premièrement, il n'y a jamais eu de base de financement solide pour l'évaluation en tant que telle dans les départements universitaires. Même lorsque l'évaluation était considérée comme une priorité majeure pour l'avenir (Wertheimer *et al.*, 1978), il y avait peu de postes ouverts dans le domaine de l'évaluation au sens strict. Les départements universitaires ont pour fonction de développer un savoir théoriquement informé et généralisable. L'évaluation poursuit rarement une telle démarche, et ne le fait que dans des circonstances particulières (généralement dans le cadre de projets fédéraux de grande envergure). De plus, dans la mesure où l'évaluation et la recherche sont considérées comme des activités distinctes, les savoirs issus de l'évaluation ont tendance, dans l'ensemble, à devenir de moins en moins généralisables. Les départements universitaires ont été et continueront d'être réticents à confier des postes permanents à des professeurs dont l'activité n'est pas orientée vers la théorie.

Deuxièmement, non seulement l'évaluation ne dispose pas de certains des attributs nécessaires à la reconnaissance académique, mais elle présente aussi d'autres caractéristiques fort peu attrayantes dans la plupart des

contextes universitaires, notamment l'exigence d'une interaction particulièrement intense avec la communauté des utilisateurs, et son caractère clairement politique. En outre, les bonnes évaluations s'appuient généralement sur de multiples domaines d'expertise, de sorte qu'avec le temps, les évaluateurs ont tendance à adopter une perspective interdisciplinaire. Étant donné que ces tendances sont probablement plus marquées dans les évaluations au niveau local qu'au niveau fédéral, et compte tenu du prestige différent de ces deux niveaux de gouvernement, les départements universitaires seront généralement mieux disposés vis-à-vis des démarches d'évaluation à l'échelon fédéral. Ceci étant, la réputation de l'évaluation dans son ensemble fait en sorte que même les projets d'évaluation relativement rigoureux peuvent souffrir d'un discrédit.

Ces facteurs se combinent pour rendre difficile l'obtention d'une titularisation pour des professeurs ou professeures de niveau débutant ayant un intérêt marqué pour l'évaluation de programmes. Il est très rare qu'une personne puisse réussir une carrière universitaire à partir de la seule démarche d'évaluation.

D'éminent-e-s chercheurs et chercheuses universitaires pratiquent toutefois l'évaluation. Ce groupe se caractérise par un haut niveau d'expertise méthodologique et/ou dans une spécialisation universitaire traditionnelle, et est souvent connu pour sa participation à des évaluations à grande échelle financées par le gouvernement fédéral, souvent des programmes fédéraux. Ce constat n'infirme pas la règle générale quant à la prévalence des critères académiques. C'est même là le modèle de l'évaluateur ou de l'évaluatrice universitaire accompli-e : soyez un véritable expert ou une véritable experte dans un domaine valorisé par votre discipline, et poursuivez des projets à grande échelle, bien financés, pour lesquels votre expertise est pertinente et qui, à leur tour, l'enrichissent. Une telle stratégie peut faire sens pour des universitaires en début de carrière; elle permet de générer de nouvelles connaissances en sciences sociales, et peut même enrichir sur certains points la pratique de l'évaluation.

Mais ce modèle d'évaluation académique a plusieurs implications importantes : tout d'abord, l'intérêt premier de cet-te universitaire sera de préserver sa crédibilité académique, et sera donc orienté vers sa thématique et sa discipline. Deuxièmement, la plupart des activités d'évaluation qui autorisent ce niveau de sophistication se dérouleront au niveau fédéral et, de façon plus marginale, au niveau des États. L'immense besoin d'évaluation des organismes locaux restera relativement délaissé. Troisièmement, en conséquence, cela fera peu avancer les connaissances et théories relevant spécifiquement de l'évaluation.

Une autre manière de formuler cette idée consiste à dire que l'évaluation peut être pratiquée en mettant l'accent soit sur la façon dont les savoirs sont utilisés, soit sur la façon dont ils permettent des montées en généralité. Il ne s'agit pas de dire qu'une de ces pratiques relèverait bien de l'évaluation et l'autre non, ou que l'une serait plus utile, voire plus vertueuse que l'autre. Simplement, leurs priorités et leurs valeurs de référence diffèrent, et ces différences ont des conséquences. Une de ces implications est que les différentes approches sont adaptées à des contextes et à des objectifs différents. Pour survivre dans leurs départements, les évaluateurs et évaluatrices académiques doivent généralement se concentrer davantage sur le potentiel de montée en généralité. Cette orientation est tendanciellement plus compatible avec les projets soutenus par le gouvernement fédéral et, dans une moindre mesure, par les États, et généralement incompatibles avec les projets locaux. Les études mettent plus l'accent sur le contenu et/ou la méthodologie que sur l'utilisation des résultats, le plus souvent limitée au domaine immédiat d'application. Au fur et à mesure que l'expérience provenant de ces études à grande échelle s'accumule, elles ont tendance à devenir de plus en plus rigoureuses et il semblerait qu'il y ait de moins en moins de raisons de les qualifier d'« évaluation » plutôt que de « recherche ».

La démarche des évaluateurs et évaluatrices est différente : ils et elles s'emploient à recueillir des données ou à rassembler les informations existantes afin de répondre aux besoins relativement clairs de leurs

clients en termes de reddition de comptes, de prise de décision ou de planification. Lorsque des études spécifiques sont effectuées, elles sont rarement rigoureuses au point de pouvoir être qualifiées de « recherche », et bien que les méthodologies changent et s'améliorent avec le temps, et que l'amélioration de la fiabilité et de la validité des informations soit un objectif, ce n'est pas ce qui guide en premier lieu la démarche d'évaluation.

Les principales tendances semblent être, d'une part, une division croissante entre l'évaluation académique et l'évaluation non académique, dans laquelle la première ressemble davantage à la recherche tandis que la seconde reste une évaluation. Deuxièmement, on observe une partition similaire entre les évaluations menées au niveau fédéral et au niveau local, les évaluations fédérales ressemblant de plus en plus à de la recherche et les évaluateurs et évaluatrices fédéraux à des chercheurs et à des chercheuses, tandis que les évaluations et les évaluateurs et évaluatrices locaux restent plus axés sur l'évaluation. Troisièmement, les universitaires pratiquant l'évaluation continueront à se préoccuper principalement de l'établissement et du maintien de leur réputation dans une thématique universitaire donnée (y compris en méthodologie), mais très peu de démontrer leur compétence en évaluation en tant que telle. Les qualifications universitaires peuvent conduire à recommander une personne pour certains types de projets d'évaluation, mais la qualification en l'évaluation ne vaudra jamais beaucoup dans un cadre universitaire.

Aucune de ces tendances ne favorise le développement d'un ensemble de théories académiques sur l'évaluation, et sans une telle base théorique, on ne peut envisager de faire de l'évaluation en tant que telle une discipline ou une spécialité académique. On peut dès lors prédire que l'implication des universitaires dans l'évaluation va avoir tendance à diminuer.

[…] Toutefois, si nous voulons considérer l'évaluation comme une discipline académique, ou une spécialité disciplinaire ou interdisciplinaire, alors l'évaluation, telle qu'elle est actuellement

pratiquée et promet de l'être à l'avenir, a un problème, à savoir qu'elle ne développe pas de fondement théorique ou empirique qui lui soit absolument spécifique.

Bien entendu, les projets d'évaluation mobilisent fréquemment des théories, mais il s'agit généralement de théories sur le sujet évalué ou de techniques méthodologiques ou statistiques spécifiques mobilisées dans le protocole d'enquête ou dans l'analyse des données. Les résultats, pour autant qu'ils soient généralisables, peuvent, à leur tour, venir alimenter des savoirs thématiques. Nous faisons appel à des domaines de recherche en sciences sociales autres que le contenu spécifique de l'évaluation afin d'améliorer la qualité du travail d'évaluation, notamment les recherches sur l'utilisation des connaissances et sur le changement organisationnel. Mais il est rare de voir des recherches menées sur l'évaluation, et encore plus rare de voir des recherches fondées sur un énoncé théorique relatif à l'évaluation en général, ou travaillant à l'élaboration d'un tel énoncé. Tant qu'une base théorique et empirique propre à l'évaluation ne sera pas développée, l'évaluation en tant que discipline ou spécialité académique distincte n'aura pas d'avenir, et sans une telle base académique, on ne peut espérer de développement théorique significatif.

L'évaluation en tant que spécialité académique

Se demander si l'évaluation a le potentiel de devenir une discipline ou une spécialité académique revient à se demander s'il existe des questions théoriques (ou des domaines de recherche) que l'évaluation pourrait définir ou revendiquer comme étant siennes. En définissant ces domaines de contenu, il ne suffit pas d'intégrer dans la pratique de l'évaluation les connaissances théoriques ou empiriques extraites de divers autres domaines. L'intégration dans la pratique est souhaitable et se fait déjà. L'enjeu est de produire des énoncés empiriques, théoriques et

généralisables sur les principes de l'évaluation, qui seraient largement applicables aux projets d'évaluation dans divers domaines. Ce dont on a besoin, c'est de recherche et d'élaboration théorique en l'évaluation.

Sur quels thèmes cette activité de recherche pourrait-elle être axée? Flaherty et Morel (1978) ont énuméré plusieurs réalisations en recherche ou en technologie qu'ils estimaient être des acquis généralisables en matière d'évaluation. Depuis lors, peu d'entre eux ont porté beaucoup de fruits. Je désignerai la rubrique « Utilisation des données et de l'information » comme la question la plus développée sur le plan de la recherche en évaluation. Si on l'interprète au sens large, il s'agit d'une question complexe, riche et variée sur le plan conceptuel. Bien que de nombreuses autres disciplines s'intéressent à ce sujet à partir de leurs cadres conceptuels spécifiques, l'utilisation des données est de toute évidence en tant que telle une question pertinente pour l'évaluation et, à en juger par le volume des articles présentés lors de conférences et le volume de revues, elle constitue déjà une préoccupation pour les évaluateurs. De nombreuses variations sur ce thème général sont possibles.

Par exemple, en termes très généraux, nous devons savoir pourquoi, ou si ou quand, l'évaluation est souhaitable. Pour ce faire, il faut aller au-delà du jugement de valeur selon lequel la recherche et les connaissances sont toujours souhaitables et déterminer comment et à quelles fins l'évaluation est utile. Cela sera probablement important au niveau de l'État et des organismes locaux, où l'on ne manque pas de groupes mobilisés pour faire valoir que les fonds consacrés à l'évaluation seraient mieux utilisés en étant directement consacrés aux services.

[…] L'utilisation semble donc être un sujet prometteur pour des développements théoriques et, à en juger par les articles de revue et les documents de conférences, elle demeure une préoccupation pour les évaluateurs de tous types dans tous les domaines de contenu. D'autres sujets sont possibles, notamment ceux qui pourraient permettre l'élaboration d'une base théorique en évaluation. Il se peut aussi qu'aucun

effort ne soit fait pour développer une théorie générale de l'évaluation, auquel cas l'évaluation aura tendance à se séparer en domaines de contenu et en factions plus ou moins orientées vers la recherche. [...] Dans de telles circonstances, l'évaluation continuerait d'être de nature athéorique, perdrait son attrait pour les collègues universitaires orientés vers la théorie et/ou continuerait à ne les intéresser que de façon secondaire par rapport à leurs thématiques de recherche principales.

Bibliographie

Flaherty, Eugenie W., et Jonathan A. Morell. 1978. « Evaluation: Manifestations of a new field ». *Journal of Evaluation and Program Planning* 1(1) : 1-10. doi : https://doi.org/10.1016/0149-7189(78)90002-2.

Wertheimer, Michael, Allan G. Barclay, Stuart W. Cook, Charles A. Kiesler, Sigmund Koch, Klaus F. Riegel, Leonard G. Rorer, Virginia L. Senders, M. Brewster Smith et Sally E. Sperling. 1978. « Psychology and the future ». *American Psychologist* 33(7) : 631-47. doi : https://doi.org/10.1037/0003-066X.33.7.631.

5. De quelques leçons durement acquises en évaluation de programme

MICHAEL SCRIVEN

[Traduit de[1] : Scriven, Michael. 1993. « Hard-Won Lessons in Program Evaluation ». *New Directions for Program Evaluation* (58). Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de John Wiley & Sons-Books.]

Le point de vue transdisciplinaire

Le nom de ce point de vue est basé sur une distinction entre les disciplines académiques classiques [...] et un ensemble de disciplines qui traitent de différents outils, méthodes et approches utilisés par les disciplines académiques classiques. Ces transdisciplines comprennent les statistiques, la mesure, la logique et, ainsi que je le suggère, l'évaluation. Chacune d'entre elles est reliée à un certain nombre de domaines appliqués. Ainsi, les statistiques sont reliées aux domaines appliqués tels que la biostatistique, la mécanique statistique et la démographie. La tâche des statistiques en tant qu'approche transdisciplinaire n'est pas l'étude des phénomènes relevant de ces domaines, mais plutôt l'étude des outils quantitatifs pour décrire ces phénomènes. Ses résultats s'appliquent à l'ensemble des disciplines qui utilisent ou devraient utiliser des statistiques, d'où le terme de transdiscipline. La logique, avec tous ses domaines d'application comme la logique des sciences sociales, est une

1. NdT : Dans ce qui précède, Scriven a décrit une série d'autres perspectives en évaluation, avant de présenter celle qu'il défend, l'évaluation comme transdiscipline.

transdiscipline extrêmement générale, mais l'évaluation est probablement la plus générale (contrairement à la logique, elle précède le langage); toutes deux sont beaucoup plus générales que la mesure ou les statistiques.

La conception de l'évaluation que je défends a trois caractéristiques fondamentales : l'une épistémologique, l'autre politique et la dernière, disciplinaire. Tout d'abord, il s'agit d'une conception objectiviste de l'évaluation. Elle défend l'idée que l'évaluation est un processus de détermination du mérite et de la valeur de programmes, de personnels ou de produits, par exemple; et qu'il s'agit-là de propriétés réelles, bien que complexes, de choses de la vie quotidienne enracinées dans un contexte complexe; et enfin qu'un degré acceptable d'objectivité et d'exhaustivité dans la recherche de ces propriétés est non seulement possible, mais souvent atteint. [...]

Deuxièmement, mon approche de l'évaluation de programmes est axée sur le consommateur ou la consommatrice plutôt que sur les enjeux de gestion. Cela ne signifie pas qu'il s'agit d'une approche de défense des consommateurs et consommatrices au sens où elle prendrait systématiquement et seulement leur parti. Simplement, elle considère leur bien-être comme la principale justification de l'existence d'un programme et accorde par conséquent à ce bien-être la même place centrale dans l'évaluation du programme en question. Cela revient notamment à rejeter une conception de l'évaluation comme aide à la décision [...], bien que l'évaluation puisse par ailleurs déboucher sur des connaissances utiles à la prise de décision. La fonction principale de l'évaluation consiste plutôt à déterminer le mérite et la valeur des programmes en fonction de l'efficacité et de l'efficience avec lesquelles ils servent les personnes touchées, en particulier celles qui reçoivent ou devraient recevoir les services fournis et celles qui financent ces programmes – généralement les contribuables ou leurs représentantes et représentants. Le bien-être et les intérêts des personnels chargés de la mise en œuvre des programmes comptent aussi; mais pour prendre un exemple, la fonction première des écoles n'est pas de fournir un emploi

aux enseignantes et aux enseignants, si bien que leur bien-être ne peut être considéré comme d'une importance comparable à celle de l'éducation des élèves.

Dans la mesure où les gestionnaires considèrent le service aux consommateurs et consommatrices comme leur objectif premier, comme ils le devraient normalement s'ils gèrent des programmes dans le secteur public ou associatif, les informations sur le mérite ou la valeur des programmes sont utiles à la décision et à la gestion [...]; et dans la mesure où les objectifs d'un programme reflètent les besoins réels des consommateurs et consommatrices, ces informations donnent une bonne indication du degré auquel le programme atteint ses objectifs [...]. Mais l'évaluation ne doit jamais présupposer que ces deux conditions sont remplies : elle doit au contraire en faire un objet d'enquête, et en pratique, ces deux hypothèses sont souvent contredites par les faits.

L'orientation de l'approche transdisciplinaire vis-à-vis des consommateurs et consommatrices nous fait franchir une étape supplémentaire dans l'établissement de la légitimité de la démarche consistant à tirer des conclusions évaluatives[2] : elle justifie la nécessité de le faire dans de nombreux cas. En d'autres termes, elle considère toute approche comme incomplète si elle ne permet pas de tirer des conclusions évaluatives. Des démonstrations pratiques de ce principe figurent dans chaque numéro du *Consumer Reports*[3] : les produits évalués sont notés et classés de manière systématique de sorte que l'on puisse voir quels sont les meilleurs du groupe (classement) et si les meilleurs sont sûrs, s'ils sont de bons achats, et ainsi de suite (notation).

2. NdT : Scriven a précédemment évoqué une tendance, dans plusieurs autres pratiques en évaluation, à mobiliser une démarche d'investigation empirique sans aller jusqu'à tirer des conclusions évaluatives.
3. NdT : revue d'une ONG états-unienne de consommateurs pratiquant des tests et évaluations de différents produits de consommation (une organisation analogue en France est l'UFC Que Choisir).

Troisièmement, l'approche adoptée ici est englobante. Elle considère l'évaluation de programmes comme l'un des nombreux domaines appliqués au sein d'une discipline globale de l'évaluation. (Ces domaines d'application peuvent correspondre à une partie d'une discipline au sens classique : par exemple l'évaluation du personnel fait partie de la psychologie industrielle et organisationnelle, la biostatistique relève de la biologie.) Cette perspective entraîne des changements importants dans l'éventail des considérations auxquelles l'évaluation de programmes doit prêter attention. Par exemple, elle doit examiner d'autres domaines d'application pour identifier des parallèles, tout en se référant à une discipline fondamentale pour les analyses théoriques. Une telle approche transversale enrichit considérablement le répertoire méthodologique de l'évaluation de programmes.

L'approche transdisciplinaire présente donc deux caractéristiques principales. La première est son caractère englobant. Celui-ci renvoie à plusieurs choses :

(1) La vaste gamme de domaines d'application de l'évaluation; par exemple, dans les sciences sociales, l'évaluation peut porter sur des programmes, des produits, du personnel, des politiques publiques, des performances, des propositions, ainsi que sur les évaluations elles-mêmes (méta-évaluation);

(2) Le large éventail de processus d'évaluation en jeu en dehors même des domaines d'application de l'évaluation, y compris dans toutes les disciplines (évaluation des méthodologies, des informations, des instruments, de la recherche, des théories, etc.), dans l'artisanat et les arts (évaluation des compétences artisanales, des compositions, des régimes, des instructions, etc.);

(3) Le large éventail de types d'enquêtes évaluatives, allant des niveaux pratiques d'évaluation (par exemple, juger de l'utilité des produits ou de la qualité des plongeons dans une compétition olympique), à l'analyse

conceptuelle (par exemple l'évaluation des questions conceptuelles et théoriques dans la discipline de base de l'évaluation), en passant par l'évaluation des programmes sur le terrain;

(4) Le chevauchement entre les différents domaines d'application de l'évaluation, qui est rarement reconnu : par exemple, les méthodes d'un domaine résolvent souvent des problèmes dans d'autres domaines. L'évaluation de programmes telle qu'elle est habituellement conçue ne fait pas référence à l'évaluation du personnel, à l'évaluation des propositions ou à l'évaluation éthique, alors qu'elle peut avoir recours à ces démarches.

La deuxième caractéristique principale de l'approche transdisciplinaire est l'accent qu'elle place sur les aspects techniques. L'évaluation, comme les statistiques et la mesure, requiert une formation à la manipulation de savoirs et de compétences spécifiques. Elle s'en distingue toutefois par son caractère beaucoup plus multidisciplinaire. Par exemple, l'évaluation de programmes implique plus d'une douzaine de sous-disciplines, dont plus de la moitié ne sont pas couvertes par les formations doctorales courantes dans des disciplines telles que la sociologie, la psychologie, le droit ou la comptabilité.

[...] Si l'on reconnaît qu'il existe une logique commune à toutes les démarches d'évaluation, que l'on peut observer dans l'évaluation de produits par exemple, il faut conclure que la discipline de l'évaluation transcende les limites de certains domaines d'application, tels que l'évaluation de programmes. [...] Même si certaines caractéristiques particulières de l'évaluation de programmes la font paraître moins évidente que la démarche plus simple d'évaluation de produits, ce n'est pas forcément le cas. L'opinion selon laquelle l'évaluation de programmes n'a rien à voir avec l'évaluation de produits n'est populaire que chez celles et ceux qui connaissent mal l'évaluation des produits. Par exemple, l'idée selon laquelle l'évaluation de programmes serait intrinsèquement beaucoup plus politique que l'évaluation de produits est courante, mais erronée; l'histoire du réseau routier interétatique et du 'supercollisionneur' 'supraconducteur' sont des contre-exemples, et c'est

après tout « seulement » une évaluation de produit – commandée par le Congrès et effectuée sans faille – qui a conduit à la révocation du directeur du *National Bureau of Standards* [Bureau national des normes].

6. L'hybridation disciplinaire, nouveau talisman de l'évaluation?

STEVE JACOB

[Traduit de : Jacob, Steve. 2008. « Cross-Disciplinarization a New Talisman for Evaluation? » *American Journal of Evaluation* 19(2) : 175–94. Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Sage Publications. Traduction relue et corrigée par l'auteur.]

Les relations entre les disciplines scientifiques prennent une diversité de formes telles que la multidisciplinarité, l'interdisciplinarité et la transdisciplinarité. Bien que ces termes soient assez vagues et aient de multiples connotations, leur dénominateur commun est leur potentielle valeur pour encourager l'échange, l'interaction et la coopération au-delà des frontières disciplinaires. Un certain nombre de chercheurs et de chercheuses ont tenté de clarifier cette question en soulignant les principales distinctions entre différents degrés d'intégration disciplinaire. Pour éviter toute querelle sémantique, nous retenons la taxonomie de Patricia L. Rosenfield (1992), considérée par beaucoup comme le point de départ le plus pertinent (Fuqua *et al.*, 2004; Maton, Perkins, Altman, *et al.*, 2006; Maton, Perkins, et Saegert, 2006; Stokols *et al.*, 2003; Sussman *et al.*, 2004). Rosenfield présente une taxonomie simple à trois niveaux de la recherche qui favorise la combinaison de différentes disciplines. Le premier type de recherche impliquant de telles collaborations, et le plus fréquent, est *multidisciplinaire* : des chercheurs et chercheuses travaillent à résoudre le même problème en parallèle et de manière séparée, à partir de perspectives disciplinaires différentes. Le deuxième niveau est *interdisciplinaire* : des chercheuses et chercheurs partant de bases disciplinaires différentes travaillent ensemble sur un problème

commun. Ils et elles utilisent leurs techniques et leurs compétences spécifiques pour développer de nouvelles connaissances qui sont ensuite rapportées de façon séquentielle, discipline par discipline, dans la présentation des résultats du projet. Le troisième niveau, la *transdisciplinarité*, est atteint lorsque « des chercheurs travaillent conjointement en utilisant des cadres conceptuels partagés rassemblant des théories, des concepts et des approches issus de différentes disciplines pour résoudre un problème commun » (Rosenfield, 1992 : 1351). À partir de ces définitions, nous pouvons conclure que l'approche transdisciplinaire a les objectifs les plus ambitieux et qu'elle pose les bases d'une métadiscipline. Une métadiscipline devient nécessaire, selon Kesteman (2004), quand on souhaite trouver « des solutions pratiques à des problèmes complexes, développer de nouvelles connaissances qui peuvent être appliquées rapidement, et prendre en compte des points de vue multiples et souvent contradictoires » (p.101).

En quoi l'hybridation disciplinaire se distingue-t-elle du développement d'une discipline nouvelle et distincte? La création d'une nouvelle discipline vise à tracer de nouvelles frontières. Les objectifs de l'hybridation disciplinaire sont ambitieux et permettent de rompre avec l'état d'autarcie disciplinaire qui caractérise encore de nombreux travaux (Dogan et Pahre, 1990). Ainsi, l'hybridation disciplinaire est « un processus qui consiste à résoudre un problème, une question ou un sujet trop vaste ou trop complexe pour être traité de manière adéquate par une seule discipline ou profession » (Klein et Newell, 1998 : 3). En outre, il s'agit d'une « nouvelle forme d'apprentissage et de résolution de problèmes impliquant une coopération entre différentes parties de la société et du monde universitaire pour relever les défis complexes de la société » (Häberli *et al.*, 2001 : 7). L'hybridation disciplinaire survient lorsqu'un groupe de chercheurs et de chercheuses décide que la croissance exponentielle des connaissances rend trop difficile l'émergence d'une vision globale du monde. Elle implique le transfert des méthodes d'une discipline à une autre sans que les chercheurs et chercheuses concerné-e-s y perdent leur affiliation disciplinaire. Il s'agit donc d'un processus

d'import-export méthodologique et conceptuel, qui dans certains cas peut donner l'impression d'une nouvelle discipline ou d'un domaine de spécialité mixte. La coopération entre les chercheurs au sein d'une même discipline et entre disciplines est un facteur clé de l'évolution des méthodes de production des connaissances.

Le cas de l'évaluation : réflexions sur la disciplinarité et la combinaison des disciplines dans un domaine transversal

Extraits p. 178-180 :

Le développement d'un nouveau domaine d'enquête et de pratique

En observant attentivement le développement de l'évaluation, on remarquera que la pratique a considérablement évolué au fil du temps. Pour être concis, il existe un consensus général sur le fait que l'objectif de l'évaluation est de juger du mérite, de la valeur et de l'importance des différents objets évalués : produits, personnel, politiques publiques, etc. Une tâche commune pour les évaluateurs et évaluatrices de programmes est par exemple l'identification et l'évaluation de l'impact des politiques publiques. Au cours des 40 dernières années, l'évaluation s'est appuyée sur le travail de chercheuses et chercheurs de différentes disciplines (de la psychologie et de la sociologie à l'économie), avant d'acquérir progressivement le statut de domaine autonome à l'égard de ces disciplines. Toutefois, un débat animé fait rage sur la question de savoir si l'évaluation est réellement une discipline distincte. Sur la base de la discussion précédente, il semblerait naturel que l'évaluation des

politiques publiques occupe une place centrale dans le débat « multi-trans-post-inter-disciplinarité », même s'il faut reconnaître que « l'évaluation en tant que discipline est une affaire très récente » (Scriven, 1994 : 147). La pratique de l'évaluation pourrait s'expliquer selon la logique du développement et de la spécialisation disciplinaire [...]. Cependant, elle est partiellement écartée de cette logique, car au départ, elle n'est pas issue d'une seule discipline. En effet, depuis sa conception, l'évaluation a emprunté son vocabulaire, ses méthodes et ses techniques à plusieurs disciplines. Néanmoins, les évaluatrices et évaluateurs, à l'instar des membres de toute autre discipline, se préoccupent de l'inévitable tendance à la spécialisation (logique interne) ainsi que du souhait permanent de coopérer avec les membres d'autres domaines (logique externe).

Être ou ne pas être une discipline

L'évaluation est-elle une discipline? Ou est-elle simplement un outil de gestion qui emprunte ses instruments à une ou plusieurs disciplines distinctes? Ces questions directes et provocatrices en soulèvent de nombreuses autres et suscitent parfois de vives discussions entre les chercheuses et chercheurs et les expert-e-s de l'évaluation. Même si l'évaluation est souvent comparée à des travaux de recherche scientifique, notamment en ce qui a trait à la rigueur méthodologique, ces deux exercices doivent être considérés comme distincts même s'ils sont parfois entrepris par les mêmes personnes.

Que l'évaluation soit considérée comme une discipline scientifique ou une pratique de gestion est un enjeu qui soulève les passions des partis en présence qui campent sur leur position avec une conviction quasi religieuse. De plus, le statut et l'importance attribués à l'évaluation varient énormément selon le point de vue que l'on adopte sur ce débat. Dans cette section, nous ne prétendons pas apporter une réponse définitive à

la question. Nous présentons plutôt, dans un premier temps, les positions divergentes sur le statut disciplinaire de l'évaluation. Le principal point de discorde dans le débat est de savoir si l'évaluation est devenue, sur le plan épistémologique, suffisamment distincte des domaines disciplinaires à partir desquels elle s'est développée pour lui permettre de revendiquer son statut de discipline distincte, ou si au contraire, ces domaines disciplinaires restent de forts déterminants pour produire des connaissances par le biais de l'évaluation. Ce débat repose sur la matrice disciplinaire structurée par Kuhn autour des généralisations symboliques, des modèles et des exemples (compris comme la solution concrète à un problème). À partir de quelques critères explicites permettant de juger du statut de discipline, nous examinons ensuite comment le domaine de l'évaluation est plus ou moins compatible avec les caractéristiques d'une discipline formelle et mature. Bien que l'évaluation partage un certain nombre de points communs avec les disciplines, nous soutenons que l'évaluation n'est peut-être pas (encore) une discipline à part entière.

Un récent débat dans *The Industrial-Organizational Psychologist* illustre la controverse sur le statut disciplinaire – ou non – de l'évaluation. L'échange a été initié par E. Jane Davidson (2002), qui affirme que l'évaluation est « une discipline à part entière » (p.33). Pour étayer son propos, elle souligne les spécificités méthodologiques développées par la profession et les compétences particulières requises pour mener une évaluation. Selon Davidson (2005), l'un des grands défis du développement de l'évaluation en tant que discipline est de la faire reconnaître comme étant distincte des diverses autres disciplines dont elle est issue (p.3). Les arguments de Davidson s'inspirent du travail de Scriven (1994), qui considère que la discipline générale de l'évaluation regroupe la somme des activités visant à apprécier la valeur ou la qualité de quelque chose (programme, personnel, performance, produit, proposition, politique, voire même l'évaluation elle-même par une méta-évaluation) (p.148). Selon Scriven, l'évaluation, comme les mathématiques, est une transdiscipline. Les transdisciplines sont un groupe d'élite de disciplines fournissant « des outils essentiels pour d'autres disciplines,

tout en conservant une structure autonome et un effort de recherche propre » (Scriven, 1993 : 19). En réponse à Davidson, Robert Perloff (2003) affirme que l'évaluation n'est pas une discipline. Il affirme que « les disciplines sont systématiques, cohérentes, fondées le plus souvent sur une théorie solide et proposées comme programmes dans les collèges, les universités et les écoles professionnelles accréditées » (p.52). D'autre part, Perloff (2003) ajoute que « l'évaluation est un mélange hétéroclite, un ragoût de procédés développés par 'essai et erreur' » (p.52).

[…]

Extraits p. 180-181 :

Nous soutenons que bien que l'évaluation soit un domaine autonome, elle est à la fois une composante et composée d'un certain nombre d'autres disciplines. Pour certain-e-s, l'avenir de l'évaluation sera ancré dans des fondements disciplinaires, mais pour d'autres, les perspectives sont transdisciplinaires (Coryn et Hattie, 2006). En même temps, le fait que l'évaluation partage un nombre d'attributs avec d'autres disciplines à part entière, tend à garantir qu'elle soit traitée comme telle. Comme nous l'avons indiqué précédemment, depuis les débuts de l'évaluation, les évaluateurs et évaluatrices ont emprunté à diverses disciplines académiques des concepts théoriques et des instruments méthodologiques nécessaires pour formuler des recommandations utiles et crédibles. À cet égard, Scriven (1994) soutient que la spécificité méthodologique de l'évaluation constitue l'élément principal qui justifie la prise en compte d'une « véritable discipline de l'évaluation » (p.148).

En ce qui concerne la pratique de l'évaluation, l'existence d'une communauté d'évaluateurs et d'évaluatrices est indéniable, comme l'indiquent clairement les nombreuses sociétés nationales et internationales dans ce domaine. Cette communauté est composée d'acteurs et d'actrices aux profils hétéroclites qui varient selon le rôle qu'ils et elles remplissent dans le processus d'évaluation (commanditaire, évaluateur ou partie prenante) et selon leur affiliation organisationnelle

(administration publique, université ou secteur privé). Si l'on ajoute à cela la multiplicité des objectifs poursuivis par la démarche d'évaluation, on peut facilement voir qu'une telle diversité rend difficile l'émergence de l'évaluation en tant que discipline spécifique. Par conséquent, il semble logique d'insister sur un ancrage disciplinaire concret de la part des évaluateurs et évaluatrices pour qu'ils et elles puissent s'acquitter de leur tâche. En reproduisant la structure institutionnelle du milieu académique qui subit des transformations et qui est régulièrement remise en question, les évaluatrices et évaluateurs reproduisent également les conditions favorisées par cette sphère encore plus segmentée.

Dès lors, c'est surtout le processus de professionnalisation de la pratique qui rend l'évaluation de plus en plus indépendante en tant que discipline (Morell, 1990; Smith, 2001). Avant d'examiner plus avant l'évaluation fondée sur l'hybridation disciplinaire pour en apprécier la valeur, soulignons d'abord brièvement certaines conséquences associées à la spécialisation de l'évaluation tendant vers le statut de discipline spécifique. À mesure qu'une discipline se spécialise, elle développe une sémantique distincte qui peut progressivement créer des barrières entre les chercheurs (Pantazidou et Nair, 2001 : 343). Pour résumer, l'évaluation a quitté les laboratoires universitaires pour la sphère privée du marché du conseil. Néanmoins, la spécialisation de l'évaluation en tant que discipline distincte n'est pas complète; les antécédents disciplinaires de l'évaluateur ou de l'évaluatrice demeurent importants et guident sa pratique. Il n'est donc pas surprenant que l'évaluation professionnelle reflète les frontières disciplinaires traditionnelles du milieu universitaire. La principale vertu découlant de la spécialisation de l'évaluation en une discipline distincte est qu'elle nous permet d'identifier et de prendre en compte ses particularités. Ces particularités découlent de la nécessité de trouver un équilibre entre une perspective académique (essentiellement pour des raisons méthodologiques) et une perspective pragmatique – nécessaire pour encourager la mise en œuvre de la recherche en évaluation par les décideurs et décideuses (Patton, 1997).

Toutefois, à l'heure actuelle, cette évolution a conduit à une spécialisation, voire à une hyperspécialisation, qui comporte elle-même certains risques. La frontière disciplinaire, avec son langage et ses concepts uniques, peut isoler les évaluatrices et évaluateurs des chercheuses et chercheurs d'autres domaines et les empêcher ainsi de se faire une idée des préoccupations ou des problèmes qu'elles et ils ont en commun. Cet esprit disciplinaire risque de propager parmi les évaluateurs et évaluatrices un air malsain d'exclusivité, qui interdirait tour empiétement extérieur sur leur domaine d'expertise (Morin, 1994). Dans un contexte d'essor des croisements disciplinaires à l'université, il est grand temps que la communauté de l'évaluation saisisse le potentiel de cette dynamique pour aborder les problèmes du monde réel.

Vers une combinaison des disciplines en évaluation

[…] Extraits p. 182-185 :

Au-delà des slogans et des déclarations d'intention, plusieurs avantages plaident en faveur d'une évaluation fondée sur l'hybridation disciplinaire.

Validité des connaissances générées. Tout d'abord, l'évaluation fondée sur l'hybridation disciplinaire permet une mise en commun des connaissances et des projections spécifiques à certains domaines. Plutôt que de considérer cette approche comme une négociation sur le tracé des frontières, il faut y voir l'émergence d'une pratique qui crée des intersections. Selon Blackwell (1955), le croisement de différentes disciplines est utile dans des situations précises. Il s'agit notamment des cas dans lesquels une discipline ne peut pas traiter un problème de manière adéquate : théoriquement, le problème se situe dans une zone grise entre les disciplines; différentes disciplines ont contribué à faire progresser la réflexion sur un problème; l'intégration de cadres conceptuels auparavant distincts semble nécessaire; le problème est d'une telle ampleur que seule la recherche en équipe pourrait le traiter;

les membres des disciplines pertinentes sont prêtes et disposées à collaborer; enfin, « des chercheurs issus des disciplines pertinentes, et qui répondent aux critères d'une recherche en équipe multidisciplinaire, sont disponibles » (Blackwell, 1955 : 370). En effet, le mouvement de spécialisation [propre aux disciplines] entraîne un affaiblissement de la communication entre les champs, voire y met fin (Spengler, 1950 : 360). C'est cette communication que nous estimons pouvoir restaurer par l'hybridation disciplinaire.

En réunissant un groupe d'expert-e-s autour d'un objet d'évaluation commun, l'hybridation disciplinaire évite une fragmentation inutile de l'objet étudié. Cette combinaison qui a pour but d'examiner les problèmes sous différents angles, peut s'avérer bénéfique et devrait permettre une synthèse qui « peut contribuer à stimuler la réflexion en dehors des sentiers battus et à perfectionner les compétences de communication entre différents secteurs d'activité et disciplines » (Davidson, 2002 : 34). Il évite ainsi la fragmentation en fonction des préoccupations spécifiques de chaque domaine disciplinaire. La réalisation d'une analyse unique, mais multidimensionnelle permet l'expression et l'intégration de points de vue contrastés. En tirant parti du potentiel de chaque discipline, ceci permet l'élaboration de théories générales intégrées ou de concepts englobants (Blackwell, 1955 : 369-70).

Utilité des conclusions. Le deuxième avantage découle en partie du premier. Il semble que c'est en matière d'utilité que les résultats d'une analyse fondée sur l'hybridation disciplinaire ont le plus de potentiel. Cette évolution modifie les critères de validité d'un résultat de recherche. L'utilité de la connaissance pour le commanditaire remplace la traditionnelle évaluation par les pairs qui était autrefois la méthode la plus courante de vérification de la validité : « le monde académique a ainsi l'occasion de corriger sa myopie qui se focalise uniquement sur des formes de connaissances scientifiques » (Muller et Subotzky, 2001 : 175). En effet, en plus de fournir des solutions à des problèmes complexes, il semble que les résultats obtenus par la combinaison des disciplines seraient plus crédibles aux yeux de certains bailleurs de fonds qui

encouragent souvent la collaboration entre praticien-ne-s et universitaires (Hackett, 2000; Stark, 1995). Ce point de vue est étayé par le constat qu'il faut combiner des perspectives multiples pour résoudre les problèmes du monde réel. En effet, « les problèmes se présentent en 'couches' qui doivent être séparées et analysées, mais les solutions doivent généralement être globales et aborder le problème comme un système, et non comme des pièces détachées » (Davis, 1995 : 39). Les connaissances issues d'une combinaison de plusieurs disciplines tiennent compte d'un certain nombre de facettes du domaine examiné, ce qui accroît la précision des résultats et renforce l'acceptabilité et la faisabilité des recommandations. En fait, l'examen de quelques exemples de recherches fondées sur un croisement de disciplines révèle que ces projets encouragent les chercheuses et chercheurs à quitter leurs laboratoires, à collaborer avec les responsables de la mise en œuvre des politiques publiques, et permettent mieux aux décideuses et décideurs d'atteindre leurs objectifs spécifiques. Enfin, la traduction des résultats en techniques appliquées peut améliorer les capacités des parties prenantes et faire émerger de nouvelles perspectives sur des interventions alternatives (Fuqua *et al.*, 2004; Maton, Perkins, et Saegert, 2006). Ainsi, les recommandations d'action issues d'études croisant les disciplines répondent au critère d'utilité, qui est pour beaucoup d'évaluateurs et d'évaluatrices l'élément le plus important pour mesurer la valeur d'une évaluation (Patton, 1997; Smith, 1979).

Capital social et apprentissage. Un autre aspect à ne pas négliger est que l'hybridation disciplinaire encourage le développement du capital social entre les personnes ainsi amenées à collaborer (chercheurs, chercheuses, évaluateurs, évaluatrices ou praticien-ne-s) qui sont associées dans un processus de construction d'un cadre commun, qui partagent des valeurs et font face ensemble aux difficultés qui se présentent (Fuqua *et al.*, 2004; Morgan *et al.*, 2003; Rosenfield, 1992; Stokols *et al.*, 2003). Cela tend à créer des communautés d'apprentissage qui encouragent le transfert de connaissances et de compétences entre les intervenant-e-s (Guthrie *et al.*, 2006). Sur ce point, l'examen systématique de la littérature sur

le transfert des connaissances dans le secteur de la santé montre que le facteur le plus souvent cité comme déterminant de l'utilisation des connaissances scientifiques réside dans les relations interpersonnelles entre les chercheurs et chercheuses d'une part, et les utilisateurs et utilisatrices d'autre part (Innvaer *et al.*, 2002; Lavis *et al.*, 2005).

Satisfaire les besoins du ou de la commanditaire. Il convient de souligner que ces arguments « théoriques » en faveur du croisement des disciplines trouvent un écho auprès des utilisatrices et utilisateurs de l'évaluation. Puisqu'elles et ils sont conscient-e-s que « de nombreux problèmes de recherche ne peuvent être facilement résolus à l'intérieur des limites de disciplines particulières » (Salter et Hearn, 1996 : 3), la plupart des commanditaires demandent une évaluation pour recueillir des résultats complets et des recommandations spécifiques. Les attentes à l'égard des évaluations sont élevées, et la plupart des commanditaires d'évaluation se préoccupent peu des frontières disciplinaires et des débats territoriaux qui structurent le monde universitaire. L'hybridation disciplinaire répond aux besoins croissants des utilisatrices et utilisateurs en matière de conseils politiques complexes et largement ciblés et pourrait améliorer les conclusions d'une évaluation de plusieurs façons. L'hybridation disciplinaire permet de dépasser les objectifs séquentiels et de mieux tenir compte d'une vision globale qui prévoit un processus décisionnel intégré au cœur de l'action, cette vision est inspirée par quelque chose ressemblant à une approche médicale holistique (Herman *et al.*, 2000; Yoshikawa, 2006).

Bénéfices de la spécialisation disciplinaire. Enfin, l'hybridation disciplinaire permet de tirer le meilleur des deux mondes, car outre les avantages mentionnés précédemment, elle bénéficie également des avancées disciplinaires propres à chaque domaine. Il est utile de noter « que toute-s les spécialistes ne sont pas forcément enclin-e-s à la coopération interdisciplinaire, ni capables de travailler de cette manière. La coopération interdisciplinaire est une tâche qui incombe à une poignée d'élu-e-s » (Wohl, 1955 : 376). Pour les évaluatrices et évaluateurs qui sont disposé-e-s à participer à cet exercice, les leçons et la formation de

chaque discipline peuvent enrichir leur façon de voir l'objet évalué. C'est grâce à un ancrage disciplinaire que les progrès sont rendus possibles et réalisables. Il faut donc souligner que le recours à une perspective qui met l'accent sur la multiplicité n'est évidemment pas une tentative d'abolir les frontières disciplinaires. « Les disciplines ne se 'fertilisent' pas mutuellement de façon naturelle comme tant de fleurs sauvages. Pour qu'un travail créatif en collaboration devienne possible, il faut prendre appui sur des groupes de chercheurs déjà reliés par des liens sociaux solides et satisfaisants » (p.376). En effet, « la collaboration interdisciplinaire découle du fait même de la spécialisation et serait inconcevable sans elle » (p.376). D'après ce qui précède, il est nécessaire de garder à l'esprit que toute discussion sur les avantages de l'évaluation fondée sur l'hybridation disciplinaire doit aussi prendre en compte les inconvénients qui pourraient en résulter.

Ou une source de problèmes insurmontables?

Une valeur ajoutée nulle en matière de validité des résultats. L'une des principales difficultés qui résulte d'une évaluation fondée sur l'hybridation disciplinaire est l'hypothèse selon laquelle la rencontre de diverses perspectives est supérieure à une approche monodisciplinaire. L'ajout de perspectives est inutile si elle ne produit pas de meilleurs résultats. Il ne faut donc pas considérer la réalisation d'une évaluation fondée sur l'hybridation disciplinaire comme une fin en soi. En outre, personne ne se satisfera d'une « dilution constante de la spécialisation sans que la synthèse ne produise des avantages en retour » (Sussman *et al.*, 2004; Wohl, 1955 : 379). L'évaluateur ou l'évaluatrice aura l'impression d'avoir perdu son temps et le ou la commanditaire aura gaspillé son argent et se trouvera avec des conclusions et des recommandations inadéquates. Les évaluateurs et les évaluatrices doivent s'assurer qu'ils et elles

comprennent vraiment les enjeux du programme qu'ils et elles évaluent avant de rendre les choses trop complexes sur le plan analytique (Johnson, 1990 : 133).

Une dégradation de la qualité des résultats de l'évaluation. La contrepartie de la complémentarité, que nous avons présentée comme l'un des avantages de la combinaison de plusieurs disciplines, est sans aucun doute le risque de produire des résultats centrés sur le plus petit dénominateur commun. Ainsi, on court le risque de « ne pas utiliser les outils et les concepts les plus sophistiqués et les plus puissants de chaque discipline lorsqu'on tente de fusionner les disciplines ou de les rendre équivalentes dans une entreprise de recherche spécifique » (Blackwell, 1955 : 370). Étant donné que « le temps et l'énergie consacrés à la compréhension d'autres disciplines nuisent invariablement au temps et à l'engagement consacrés à la maximisation de la maîtrise d'une seule discipline » (Naiman, 1999 : 293), certain-e-s considèrent que les approches fondées sur l'hybridation disciplinaire compromettent l'ensemble du processus analytique. Il y a donc un risque d'être non scientifique (Caldwell, 1996). Il faut par conséquent éviter d'importer certaines théories d'une discipline donnée ou de surestimer l'utilité de nouvelles connaissances ou théories avant que leur validité n'ait été rigoureusement testée pour en assurer la validité. Il pourrait aussi arriver que la pensée complexe et la réflexion pointue cèdent la place à une sorte de surf scientifique et intellectuel. Cette inclinaison à la superficialité sape rapidement le développement de connaissances spécialisées (Hamel, 2005 : 109).

Des rapports tendus entre les disciplines. Les approches fondées sur l'hybridation disciplinaire pourraient également se heurter à une sorte de chauvinisme disciplinaire de la part de ceux qui ont une faible estime quant à cette façon de travailler, ou à une sorte d'impérialisme disciplinaire de la part de ceux qui pensent que leur discipline est supérieure aux autres (Younglove-Webb *et al.*, 1999). Il existe encore une croyance selon laquelle « aucun prestige n'est accordé à ceux qui travaillent avec d'autres disciplines » (Rosenfield, 1992 : 1355). Cet

impérialisme disciplinaire est miné par les travaux fondés sur l'hybridation disciplinaire qui ont le potentiel de redéfinir de nouvelles bases, de remettre en question les points de vue traditionnels et d'aboutir à de nouveaux domaines de recherche. Entre-temps, il ne faut pas oublier que l'hybridation des disciplines émerge dans un contexte où les disciplines conservent un rôle essentiel, car, comme le suggère Friedman (2001), « Peut-on avoir de l'interdisciplinarité sans qu'il y ait un certain sentiment de dépassement ou de transgression des frontières disciplinaires? » (p.506).

Une disjonction entre les niveaux d'analyse. Une autre menace qui pèse sur ce type de travail est le risque d'inintelligibilité en raison d'une disjonction quasi totale des niveaux d'analyse. En outre, l'avantage attendu de la communication entre des domaines de connaissances dissociés oblige à sacrifier certaines de ses valeurs pour prendre en compte les « faits » ou les valeurs d'un autre domaine. Cela dépasse les difficultés sémantiques qui peuvent être surmontées assez facilement en accordant une attention particulière au partage des connaissances et au développement d'un vocabulaire commun. Ici, il ne faut pas perdre de vue qu'en dépit de tous les efforts déployés pour « unifier » la compréhension et développer des concepts communs utiles, les disciplines restent dominantes dans l'appréhension des problèmes, dans leur formulation et leur résolution. Dans ce cas, « le caractère multidisciplinaire de l'évaluation isolerait les débats théoriques et empiriques dans les disciplines traditionnelles » (Dubois et Marceau, 2005 : 25).

Des problèmes de faisabilité et de coordination. Enfin, il ne faut pas laisser croire que cette approche est plus facile à mettre en œuvre. Au contraire, si de grands efforts d'organisation ne sont pas déployés, l'évaluation fondée sur l'hybridation disciplinaire posera de nouveaux problèmes en ce qui a trait à la collaboration et à la motivation des participants. Il existe un réel risque de blocage qui pourrait affecter considérablement le processus d'évaluation. Chacun sait que le temps est une ressource inestimable dans la conduite de l'évaluation et qu'il serait donc préjudiciable de le gaspiller dans des problèmes de coordination. Il est

donc important d'accorder une plus grande attention aux compétences requises pour le travail de groupe afin que les efforts de collaboration puissent porter leurs fruits. Le travail d'équipe et l'intégration par la combinaison de disciplines différentes exigent donc une réflexion préalable sur les modalités de sa mise en œuvre afin que la collaboration soit à la fois viable et rentable (Morgan *et al.*, 2003; Suarez-Balcazar *et al.*, 2006).

Bibliographie

Blackwell, Gordon W. 1955. « Multidisciplinary team research ». *Social Forces* 33(4) : 367-74. doi : https://doi.org/10.2307/2573009.

Caldwell, John C. 1996. « Demography and social science ». *Population Studies* 50(3) : 305-33. doi : https://doi.org/10.1080/0032472031000149516.

Coryn, Chris L., et John A. Hattie. 2006. « The transdisciplinary model of evaluation ». *Journal of MultiDisciplinary Evaluation* 3(4) : 107-14.

Davidson, E. Jane. 2002. « The discipline of evaluation: A helicopter tour for I-O psychologists ». *Industrial-Organizational Psychologist* 40(2) : 31-35.

Davidson, E. Jane. 2005. « Marketing evaluation as a profession and a discipline ». *Journal of MultiDisciplinary Evaluation* 2 : 3-10.

Davis, James R. 1995. *Interdisciplinary courses and team teaching: New arrangements for learning*. Phoenix: American Council on Education and the Oryx Press.

Dogan, Mattei, et Robert Pahre. 1990. *Creative marginality: Innovation at the intersections of social sciences*. Boulder: Westview Press.

Dubois, Nathalie, et Richard Marceau. 2005. « Un état des lieux théoriques de l'évaluation : Une discipline à la remorque d'une révolution scientifique qui n'en finit pas [A theoretical inventory of evaluation: A discipline towed along by a scientific revolution that fails to complete it] ». *Revue canadienne d'évaluation de programme* 20(1) : 1-36.

Friedman, Susan S. 2001. « Statement: Academic feminism and interdisciplinarity ». *Feminist Studies* 27(2) : 504-9. doi : https://doi.org/10.2307/3178774.

Fuqua, Juliana, Daniel Stokols, Jennifer Gress, Kimari Phillips et Richard Harvey. 2004. « Transdisciplinary collaboration as a basis for enhancing the science and prevention of substance use and "abuse." » *Substance Use & Misuse* 39 : 1457-1514. doi : 10.1081/LSUM-200033200.

Guthrie, Jill, Phyll Dance, Carmen Cubillo, David McDonald, Julie Tongs, Tom Brideson et Gabriele Bammer. 2006. « Working in partnership: Skills transfer in developing a cross-cultural research team ». *Journal of Community Psychology* 34(5) : 515-22. doi : 10.1002/jcop.20112.

Häberli, Rudolf, Alain Bill, Walter Grossenbacher-Mansuy, Julie T. Klein, Roland W. Scholz et Myrtha Welti. 2001. « Synthesis ». in *Transdisciplinarity: Joint problem solving among science, technology, and society. An effective way for managing complexity*, édité par J. T. Klein, W. Grossenbacher-Mansuy, R. Häberli, A. Bill, R. W. Scholz et M. Welti. Basel Switzerland: Birkhäuser Verlag, p. 6-22.

Hackett, Edward J. 2000. « Interdisciplinary research initiatives at the U.S. National Science Foundation ». in *Practising interdisciplinarity*, édité par P. Weingart et N. Stehr. Toronto: University of Toronto Press, p. 248-59.

Hamel, Jacques. 2005. « Sociologie et interdisciplinarité, un mariage de raison [Sociology and interdisciplinarity: A marriage of convenience?] ». *A Contrario* 3(1) : 107-15. doi : 10.3917/aco.031.0107.

Herman, Sandra E., Kenneth A. Frank, Carol T. Mowbray, Kurt M. Ribisl et William S. Davidso. 2000. « Longitudinal effects of integrated treatment on alcohol use for persons with serious mental illness and substance use disorders ». *Journal of Behavioral Health Services and Research* 27:286-302. doi : https://doi.org/10.1007/BF02291740.

Innvaer, Simon, Gunn Vist, Mari Trommald et Andrew Oxman. 2002. « Health policy-makers' perceptions of their use of evidence: A systematic review ». *Journal of Health Services Research & Policy* 7(4) : 239-44. doi : 10.1258/135581902320432778.

Johnson, Paul L. 1990. « Ron Carlson and James Bell discuss the challenges facing health evaluators in the 1990s ». *American Journal of Evaluation* 11(2) : 127-33. doi : https://doi.org/10.1177/109821409001100206.

Kesteman, Jean-Pierre. 2004. « L'un, le multiple et le complexe. L'université et la transdisciplinarité [The one, the multiple and the complex: The university and transdisciplinarity] ». *A Contrario* 2(1) : 89-108. doi : 10.3917/aco.021.108.

Klein, Julie T., et William H. Newell. 1998. « Advancing interdisciplinary studies ». in *Interdisciplinarity: Essays from the literature*, édité par W. H. Newell. New York: College Entrance Examination Board, p. 3-22.

Lavis, John, Huw Davies, Andy Oxman, Jean-Louis Denis, Karen Golden-Biddle et Ewan Ferlie. 2005. « Towards systematic reviews that inform health care management and policy-making ». *Journal of Health Services Research & Policy* 10 : 35-48. doi : 10.1258/1355819054308549.

Maton, Kenneth, Douglas D. Perkins, D. G. Altman, Lorraine Gutierrez, James G. Kelly, Julian Rappaport et Susan Saegert. 2006. « Community-based interdisciplinary research: Introduction to the special issue ». *American Journal of Community Psychology* 38 : 1-7. doi : 10.1007/s10464-006-9063-2.

Maton, Kenneth, Douglas D. Perkins, et Susan Saegert. 2006. « Community psychology at the crossroads: Prospects for interdisciplinary research ». *American Journal of Community Psychology* 38(1-2) : 9-21. doi : 10.1007/s10464-006-9062-3.

Morell, Jonathan A. 1990. « Evaluation: Status of a loose coalition ». *Evaluation Practice* 11(3) : 213-19. doi : https://doi.org/10.1177/109821409001100306.

Morgan, Glen D., Kimberly Kobus, Karen K. Gerlach, Charles Neighbors, Caryn Lerman, David B. Abrams et Barbara Rimer. 2003. « Facilitating transdisciplinary research: The experience of the transdisciplinary tobacco use research centers ». *Nicotine & Tobacco Research* 5 (Suppl.1) : 11-19. doi : https://doi.org/10.1080/14622200310001625537.

Morin, Edgar. 1994. « Interdisciplinarité et transdisciplinarité ». *Transversales, Science, Culture* 29 : 4-8.

Muller, Johan, et George Subotzky. 2001. « What knowledge is needed in the new millennium ». *Organization* 8(2) : 163-82. doi : 10.1177/1350508401082004.

Naiman, Robert J. 1999. « A perspective on interdisciplinary science ». *Ecosystems* 2(4) : 292-95.

Pantazidou, M., et Indira Nair. 2001. « Academe vs Babel ». *Organization* 8(2) : 341-48. doi : 10.1177/1350508401082017.

Patton, Michael Q. 1997. *Utilization-focused evaluation: The new century text*. 3e éd. Thousand Oaks: Sage Publications.

Perloff, Robert. 2003. « A potpourri of cursory thoughts on evaluation ». *Industrial-Organizational Psychologist* 40(3) : 52-54.

Rosenfield, Patricia L. 1992. « The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences ». *Social Sciences & Medicine* 35 : 1343-57. doi : https://doi.org/10.1016/0277-9536(92)90038-R.

Salter, Liora, et Alison M. V. Hearn. 1996. *Outside the lines: Issues in interdisciplinary research*. Montreal : McGill-Queen's University Press.

Scriven, Michael. 1993. *Hard-Won Lessons in Program Evaluation*. San Francisco: Jossey-Bass.

Scriven, Michael. 1994. « The final synthesis ». *Evaluation Practice* 15(3) : 367-82. doi : 10.1177/109821409401500317.

Smith, M. F. 2001. « Evaluation: Preview of the future #2 ». *American Journal of Evaluation* 22(3) : 281-300. doi : 10.1177/109821400102200302.

Smith, Nick L. 1979. « Evaluation reflections: Requirements for a discipline of evaluation ». *Studies in Educational Evaluation* 5 : 5-12.

Spengler, Joseph J. 1950. « Generalists versus specialists in social science: An economist view ». *American Political Science Review* 44(2) : 258-379. doi :10.2307/1950276.

Stark, C. Robert. 1995. « Adopting multidisciplinary approaches to sustainable agriculture research: Potentials and pitfalls ». *American Journal of Alternative Agriculture* 10(4) : 180-83. doi : 10.1017/S0889189300006445.

Stokols, Daniel, Juliana Fuqua, Jennifer Gress, Richard Harvey, Kimari Phillips, Lourdes Baezconde-Garbanati, Jennifer Unger, Paula Palmer, Melissa A. Clark, Suzanne M. Colby, Glen Morgan et William Trochim. 2003. « Evaluating transdisciplinary science ». *Nicotine & Tobacco Research* 5 (Suppl.1) : 21-39. doi : 10.1080/1462220031000162555.

Suarez-Balcazar, Yolanda, Maureen Hellwig, Joanne Kouba, La Donna Redmond, Louise Martinez, Daniel Block, Claire Kohrman et William Peterman. 2006. « The making of an interdisciplinary partnership: The case of the Chicago food system collaborative ». *American Journal of Community Psychology* 38(1-2) : 113-23. doi : 10.1007/s10464-006-9067-y.

Sussman, Steve, Alan W. Stacy, Anderson C. Johnson, Mary Ann Pentz et Elizabeth Robertson. 2004. « A transdisciplinary focus on drug abuse prevention: An introduction ». *Substance Use & Misuse* 39(10-12) : 1441-56. doi : https://doi.org/10.1081/JA-200033194.

Wohl, Richard R. 1955. « Some observations on the social organization of interdisciplinary social science research ». *Social Forces* 33(4) : 374-83. doi : https://doi.org/10.2307/2573010.

Yoshikawa, Hirokazu. 2006. « Placing community psychology in the context of the social, health and educational sciences: Directions for interdisciplinary research and action ». *American Journal of Community Psychology* 38(1-2) : 31-34. doi : 10.1007/s10464-006-9068-x.

Younglove-Webb, Julie, Barbara Gray, Charles W. Abdalla, et Amy P. Thurow. 1999. « The dynamics of multidisciplinary research teams in academia ». *Review of Higher Education* 22(4) : 425-40.

7. Qu'est-ce que l'évaluation? En quoi diffère-t-elle (ou non) de la recherche?

DANA WANZER

[Traduit de : Wanzer, Dana. 2020. « What Is Evaluation? Perspectives of How Evaluation Differs (or Not) From Research ». *American Journal of Evaluation*. Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

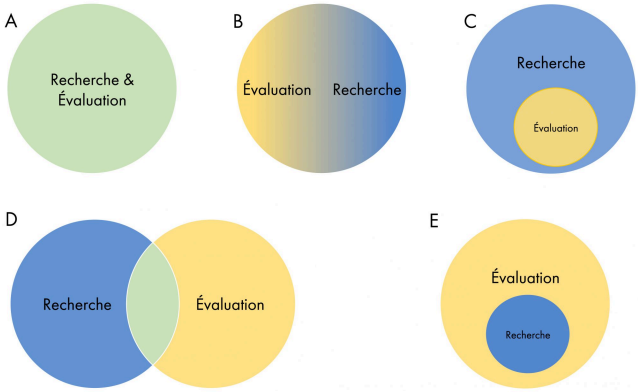
Différentes façons de distinguer (ou pas) l'évaluation de la recherche ont été proposées (voir la figure 1), et ce débat « hante le domaine » (Patton 2008 : 40). Certain-e-s affirment que l'évaluation est de la recherche et qu'il n'y a pas ou peu de distinctions entre les deux (figure 1A). Par exemple, certain-e-s affirment explicitement que « l'évaluation est une recherche appliquée… » (Barker, Pistrang, et Elliott, 2016; Hackbarth et Gall, 2005; Rallis, 2014), tout en ajoutant parfois des réserves. Toutefois, nombreuses et nombreux sont celles et ceux qui reconnaissent au moins certaines différences entre l'évaluation et la recherche. Il y a quatre façons principales de les distinguer : l'évaluation et la recherche comme les deux extrêmes d'un *continuum* (figure 1B), l'évaluation comme sous-ensemble de la recherche (figure 1C), l'évaluation et la recherche se chevauchant selon un diagramme de Venn (figure 1D) et la recherche comme sous-ensemble de l'évaluation (figure 1E). Bien que Mathison (2008) défende une vision de la recherche et de l'évaluation comme les deux extrêmes d'un *continuum* (figure 1B), cette représentation ne semble pas si courante, puisque nous n'avons trouvé aucune référence bibliographique à l'appui de cette approche. La figure 1C considère que l'évaluation n'est qu'un outil méthodologique dans la boîte à outils de la recherche (Greene, 2016), cet outil étant souvent résumé à la méthode expérimentale par assignation aléatoire. Par exemple, comme l'affirme

un évaluateur, « la recherche a une portée plus large que la recherche évaluative » (Vedung, 2004 : 111). Pour Guenther et Arnott (2011), dans le domaine de l'éducation, la « distinction tacite entre la recherche et l'évaluation fait de la seconde une composante de la première » (p.11).

Finalement, la représentation de la recherche et de l'évaluation dans un chevauchement partiel (diagramme de Venn, figure 1D) est probablement la plus courante, en particulier parmi les évaluatrices et évaluateurs (Lavelle, 2010; Mertens, 2014; Russ-Eft et Preskill, 2009; Vedung, 2004). Selon ce point de vue, il existe des similitudes entre la recherche et l'évaluation, notamment en ce qui concerne les conceptions et les méthodes utilisées, mais aussi de nombreuses différences. Une différence principale réside dans le fait que les évaluateurs et évaluatrices « posent et répondent explicitement à des questions évaluatives (c'est-à-dire des questions sur la qualité, la valeur et l'importance) » (Davidson, 2014 : 37). L'évaluation emprunte fortement aux sciences sociales en ce qui concerne les méthodes et les conceptions utilisées : ce sont là ses similitudes avec la recherche. Mais on peut par ailleurs distinguer l'évaluation et la recherche de nombreuses manières, notamment en ce qui concerne les objectifs et les résultats. Le tableau 2 décrit quelques-unes des principales différences entre l'évaluation et la recherche. Ces distinctions peuvent apparaître comme des simplifications ou des généralisations excessives (Mathison, 2008). Par exemple, certaines évaluations sont effectivement généralisables et publiées, et certaines recherches sont menées en collaboration avec des praticiens et, par conséquent, ces derniers en sont le principal auditoire. En considérant l'évaluation et la recherche comme se chevauchant, on reconnaît qu'elles présentent des similitudes mais qu'elles se développent également en parallèle. Par conséquent, tout comme la recherche est généralement considérée comme un moyen d'informer l'évaluation, l'évaluation est tout aussi à même d'éclairer la recherche (Mertens, 2014). De plus, tout comme nous faisons de la recherche sur l'évaluation, nous évaluons aussi souvent la recherche (par exemple, par le biais de publications dans des revues évaluées par des pairs, ou la méta-recherche).

Une dernière représentation possible considère la recherche comme un sous-ensemble de l'évaluation (figure 1E). L'évaluation est alors considérée comme une transdiscipline qui fournit des outils (par exemple, la pensée évaluative) à d'autres disciplines (par exemple, la recherche dans plusieurs domaines), tout en restant une discipline autonome à part entière (Scriven, 2008). Toutefois, envisager l'évaluation comme transdiscipline suppose déjà qu'elle soit une discipline (avec, par exemple, des objectifs communs, des normes de pratique, des forums professionnels, et une structure disciplinaire; (Montrosse-Moorhead, Bellara, et Gambino, 2017). Certain-e-s affirment que l'évaluation est une discipline et une profession (Donaldson et Christie, 2006; Montrosse-Moorhead *et al.*, 2017; Morris, 2007; Picciotto, 2011). Pour ces dernier-e-s, cette opinion se justifie parce qu'il existe (a) des programmes de formation et que le domaine possède des connaissances spécialisées, (b) des principes éthiques codifiés, (c) une autonomie liée à des connaissances spécialisées et des qualifications formelles spécifique, et (d) une auto-discipline. Cependant, d'autres récusent cette idée que l'évaluation soit une discipline, une profession ou les deux (Picciotto, 2011; Rossi, Lipsey, et Freeman, 2004). En outre, à l'origine, Scriven (2016) a proposé l'idée de l'évaluation en tant que transdiscipline et a présenté sa liste « Quelque chose de plus » [*Something More*] qui décrit l'évaluation comme quelque chose de plus que la recherche (Scriven, 2003). Cependant, il a également fait valoir que la recherche et l'évaluation « se chevauchent massivement », mais « qu'il y a quelques différences » (Scriven, 2016 : 33), ce qui suggère peut-être que la définition du diagramme de Venn est la façon la plus appropriée de caractériser la recherche et l'évaluation plutôt que la recherche en tant que sous-ensemble de l'évaluation.

Figure 1 : Cinq relations possibles entre l'évaluation et la recherche



Bibliographie

Barker, Chris, Nancy Pistrang et Robert Elliott. 2016. *Research methods in clinical psychology: An introduction for students and practitioners*. 3e éd. Glasgow: Wiley-Blackwell.

Davidson, E. Jane. 2014. « How "beauty" can bring truth and justice to life ». *New Directions for Evaluation* 2014(142) : 31-43.

Donaldson, Stewart I., et Christina A. Christie. 2006. « Emerging career opportunities in the transdiscipline of evaluation science ». P. 243-59 in *Applied psychology: New frontiers and rewarding careers*, édité par S. I. Donald, D. E. Berger et K. Pezdek. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Greene, Jennifer C. 2016. « Advancing equity: Cultivating an evaluation habit ». in *Evaluation for an equitable society*, édité par S. I. Donaldson et R. Picciotto. Charlotte, NC: Information Age Publishing, p. 49-66.

Guenther, John, et Allan Arnott. 2011. « Legitimising evaluation for vocational learning: From bastard sibling to equal brother ». in *AVETRA 14th Annual Conference*. Melbourne.

Hackbarth, Diana, et Gail B. Gall. 2005. « Evaluation of school-based health center programs and services: The whys and hows of demonstrating program effectiveness ». *The Nursing Clinics of North America* 40(4) : 711-24. doi : 10.1016/j.cnur.2005.07.008.

Lavelle, John M. 2010. « Describing evaluation ». AEA365. Consulté (https://aea365.org/blog/john-lavelle-on-describing-evaluation/).

Mathison, Sandra. 2008. « What Is the Differences between Evaluation and Research – and Why Do We Care? » in *Fundamental Issues in Evaluation*, édité par N. L. Smith et P. R. Brandon. New York: Guilford Press, p. 188-95.

Mertens, Donna M. 2014. *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. 4e éd. Thousand Oaks: Sage Publications.

Montrosse-Moorhead, Bianca, Aarti P. Bellara et Anthony J. Gambino. 2017. « Communicating about evaluation: A conceptual model and case example ». *Journal of Multidisciplinary Evaluation* 13(29) : 16-30.

Morris, Michael. 2007. « Ethics and evaluation ». in *Evaluation ethics for best practice: Cases and commentaries*, édité par M. Morris. New York: The Guilford Press, p. 230.

Patton, Michael Q. 2008. *Utilization-focused evaluation*. 4e éd. Los Angeles: Sage Publications.

Picciotto, Robert. 2011. « The logic of evaluation professionalism ». *Evaluation* 17(2) : 165-80. doi : https://doi.org/10.1177/1356389011403362.

Rallis, Sharon F. 2014. « When and how qualitative methods provide credible and actionable evidence: Reasoning with rigor, probity, and transparence ». in *Credible and actionable evidence: The foundation for rigorous and influential evaluations*, édité par C. A. Christie, M. M. Mark et S. L. Donaldson. Thousand Oaks: Sage Publications, p. 137-56.

Rossi, Peter H., Mark W. Lipsey et Howard E. Freeman. 2004. *Evaluation: A systematic approach*. 7e éd. Thousand Oaks: Sage Publications.

Russ-Eft, Darlene, et Hallie Preskill. 2009. *Evaluation in organizations: A systematic approach to enhancing learning, performance, and change*. New York: Basic Books.

Scriven, Michael. 2003. « Evaluation in the new millennium: The transdisciplinary vision ». in *Evaluating social programs and problems: Visions for the new millennium*, édité par S. I. Donaldson et M. Scriven. London: Routledge, p. 19-41.

Scriven, Michael. 2008. « The concept of a transdiscipline: And of evaluation as a transdiscipline ». *Journal of Multidisciplinary Evaluation*, 5(10) : 65-66.

Scriven, Michael. 2016. « Roadblocks to recognition and revolution ». *American Journal of Evaluation*, 37(1) : 27-44. doi : https://doi.org/10.1177/1098214015617847.

Vedung, E. 2004. « Evaluation research and fundamental research ». in *Evaluationsforschung : Grundlagen und ausgewählte Forschungsfelder [Evaluation research: Basics and selected fields of research]*, édité par R. Stockmann. Opladen : Leske + Budrich, p. 111-34.

8. La science de l'évaluation

MICHAEL Q. PATTON

[Traduit de : Patton, Michael Quinn. 2018. « Evaluation Science ». *American Journal of Evaluation* 39(2) : 183–200. Traduction par Carine Gazier et Anne Revillard; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

[Interrogé sur les raisons de sa participation à la Marche pour la Science de 2017, Michael Patton avait spontanément répondu : « Je suis un scientifique en évaluation. Je fais de la science de l'évaluation ». Cette anecdote sert de point de départ à une réflexion sur l'importance de revendiquer le statut scientifique de l'évaluation dans un contexte d'attaques politiques contre la démarche scientifique.]

Tant la science en général que l'évaluation en particulier sont des processus fondés sur des données factuelles, dont les conclusions sont tirées d'enquêtes systématiques visant à comprendre et à expliquer le fonctionnement de certains aspects du monde. De nos jours, la crédibilité des preuves scientifiques fait l'objet de nombreuses attaques. Coupable par association, c'est aussi la crédibilité des preuves en évaluation qui s'en trouve atteinte. Défendre la valeur des preuves scientifiques, c'est donc aussi défendre la valeur des preuves en évaluation. Il est de notre intérêt, en tant qu'évaluateurs et évaluatrices, de faire cause commune avec celles et ceux qui soutiennent la science. Les barbares antisciences ne sont pas seulement à nos portes, ils et elles sont entré-e-s et ont pris le contrôle du château. C'est le constat qui a motivé la Marche pour la Science. […] Sur les plans culturel et politique, les tendances antiscientifiques comprennent les « faits alternatifs », les « fausses nouvelles » et l'idée d'un monde « post-vérité ». En novembre 2016, les Dictionnaires Oxford ont annoncé la post-vérité comme le mot international de l'année, en

en proposant la définition suivante : « Se rapporte ou dénote les circonstances dans lesquelles des faits objectifs influencent moins l'opinion publique que les appels à l'émotion et aux convictions personnelles ».

[…] Dans une culture politique post-vérité, la science devient une perspective comme une autre. Les preuves scientifiques ne valent alors pas plus que l'opinion personnelle. Des groupes politiques de différentes appartenances ne se contentent plus de prôner des valeurs, mais promulguent leurs propres « faits ». La distinction entre la preuve et l'opinion se brouille. Par extension, cela a un effet corrosif sur l'évaluation et la délégitime. Les conclusions de l'évaluation ne deviennent qu'un autre type d'opinion. La vérité (Blackburn, 2005), qu'il s'agisse de la découvrir ou de l'exprimer, est dévalorisée et contestée.

Au-delà des affirmations selon lesquelles le changement climatique est un canular et des doutes sur la sécurité des vaccins, les tendances antiscientifiques induisent des réductions budgétaires pour la recherche scientifique, y compris des organismes comme les Instituts nationaux de la santé et l'Agence de protection de l'environnement.

La science de l'évaluation

[…] Le Conseil de la Science (*Science Council*) définit la science comme « la poursuite et l'application des connaissances et de la compréhension du monde naturel et social selon une méthodologie systématique fondée sur des preuves »[1]. Les données et les preuves sont également au fondement de la pratique de l'évaluation. […]

1. http://sciencecouncil.org/about-us/our-definition-of-science/.

Une définition de la science de l'évaluation

[…] La science consiste à étudier de façon systématique comment le monde fonctionne. La science de l'évaluation consiste à étudier de façon systématique comment, et avec quel succès, des interventions visant à changer le monde fonctionnent. La science de l'évaluation implique une investigation systématique sur le mérite, la valeur, l'utilité et l'importance de tout ce qui est évalué, en se conformant à des normes scientifiques qui comprennent l'utilisation de la logique, l'utilisation de méthodes transparentes, la soumission des résultats à une vérification, ainsi que la fourniture de preuves et de justifications explicites à l'appui de l'interprétation, de l'établissement de la valeur et du jugement. […]

La science de l'évaluation en tant que corpus disciplinaire de connaissances

Les différentes disciplines scientifiques se distinguent par les questions générales qu'elles posent et par le corpus de connaissances qui se développe en réponse à ces questions. L'évaluation en tant que spécialisation scientifique pose les questions suivantes : *Quels sont les facteurs qui contribuent à la réussite ou à l'échec des interventions? Quelles méthodes permettent de les déterminer? Quels sont les critères de jugement de la réussite ou de l'échec?* Les interventions sont tout effort, programme, projet, initiative, produit, politique, organisation ou développement communautaire, ou toute activité visant à susciter le changement. Une évaluatrice ou un évaluateur expert-e sait comment procéder à l'évaluation d'une intervention particulière dans un contexte particulier et dans un but précis. En tant que spécialiste dans la *discipline de l'évaluation*, un évaluateur ou une évaluatrice compétent-e contribue et a accès à un *corpus de connaissances* sur les façons d'étudier et de juger les interventions, et sur les façons d'appliquer les connaissances pour

concevoir et améliorer celles-ci, sur la base de modèles de réussite validés à la fois empiriquement et théoriquement par les évaluations de nombreuses autres interventions. C'est parce que l'évaluation est devenue *un réservoir de connaissances sur l'efficacité* que nous sommes consultés sur la manière de concevoir, de planifier et de mettre en œuvre de nouvelles interventions, et pas seulement pour les évaluer une fois qu'elles sont mises en œuvre.

Dans leur classique *Foundations of Program Evaluation: Theories of Practice* [Fondements de l'évaluation des programmes : Théories de la pratique], Shadish, Cook et Leviton (1991) ont examiné l'évaluation en tant que spécialité méthodologique et en tant que pratique professionnelle, mais c'est bien la théorie de l'évaluation, rendue cohérente par un corpus de connaissances validées, qui a fait de l'évaluation une discipline scientifique.

Les évaluateurs et évaluatrices de programmes développent progressivement un ensemble unique de connaissances qui distingue l'évaluation des autres spécialités tout en justifiant de les compter parmi celles-ci. L'évaluation est diverse à bien des égards, mais son potentiel d'unité intellectuelle est « *la logique de l'évaluation* », ce qui pourrait permettre de surmonter les frontières disciplinaires qui séparent les évaluateurs (p.31).

[…] Michael Scriven, philosophe et penseur pionnier de l'évaluation, distingue quatre critères permettant de reconnaître une discipline scientifique :

1. Un sujet distinct
2. Des méthodes spécifiques relatives à l'étude de ce sujet
3. Un domaine d'application important, et
4. L'obtention de résultats contribuant à une amélioration sociale et intellectuelle substantielle

Il estime que l'évaluation répond à tous ces critères (Scriven, 2004 : 186). […] Scriven a été le premier à considérer l'évaluation comme la science de l'attribution de la valeur (Shadish *et al.*, 1991 : 74). Selon ses termes, « une grande partie de mes premiers travaux en évaluation avait pour enjeu de contrer diverses tentatives visant à marginaliser l'évaluation en la renvoyant dans la catégorie de l'aide à la décision, *par opposition* à la démarche de la « vraie science » recherchant la vérité » (Scriven, 2004 : 188). Pour Scriven, l'évaluation est sans aucun doute « *scientifiquement légitime*… L'essentiel… est qu'un évaluateur ou une évaluatrice de programme compétent-e puisse démontrer *scientifiquement*, par exemple, qu'un programme d'enseignement de la lecture est vraiment excellent ou vraiment sans valeur » (Scriven, 2013 : 171).

Ici, comme pour le statut professionnel de l'évaluation, les universitaires débattront entre eux de la question de savoir si, et dans quelle mesure, l'évaluation est une discipline scientifique. Se disputer sur les définitions et les limites, c'est ce que font les universitaires, et j'ai eu ma part de plaisir à le faire. Mais, dans le monde extérieur, lorsque je parle d'évaluation, je proclame, j'affirme et je célèbre sans réserve le statut disciplinaire de l'évaluation en tant que corpus de connaissances scientifiques disposant de ses propres fondements théoriques. Nos revues évaluées par des pairs réunissent, évaluent et diffusent cet ensemble de connaissances. La nature et la profondeur de notre statut disciplinaire peuvent faire l'objet d'un débat universitaire, mais le fait que nous disposons d'un réservoir de théories et de connaissances scientifiques devrait, selon moi, faire partie de notre personnalité et de notre identité publique vis-à-vis du monde entier.

La science de l'évaluation en tant que transdiscipline

Plus récemment, Scriven (2008) a envisagé et préconisé avec éloquence et force de positionner l'évaluation comme la *transdiscipline alpha* : en tant que transdiscipline, l'évaluation est au sommet de la hiérarchie disciplinaire, scientifique, académique et scientifique, avec la philosophie, la logique et la statistique, comme corpus de connaissances, de théories et de méthodes qui sont essentiels à l'érudition, à la création de connaissances et à la rigueur scientifique de toutes les autres disciplines. [...] Scriven n'aspire pas à ce que l'évaluation devienne davantage une science, car il a longtemps critiqué le fait que de nombreuses pratiques scientifiques négligent l'importance du jugement (Scriven, 1976). Sa démarche consiste plutôt à mettre la science au défi de devenir plus évaluative. [...]

Autres fondements de l'évaluation en tant que science

Le livre de Stewart Donaldson (2007) sur *Program Theory-Driven Evaluation Science* [Une science de l'évaluation basée sur la théorie des programmes] a défini le terme comme suit :

> La science de l'évaluation (plutôt que l'évaluation) vise à mettre en exergue l'utilisation de méthodes scientifiques rigoureuses (c'est-à-dire des méthodes qualitatives, quantitatives et mixtes) pour tenter de répondre à des questions valorisées dans le domaine de l'évaluation. Dans la pratique de l'évaluation, il est particulièrement important d'avoir recours à des méthodes scientifiques systématiques pour surmonter la réputation négative de la profession dans certains contextes. En d'autres termes, dans certains contextes, l'évaluation est critiquée pour son manque de fiabilité, sa mollesse ou son caractère de second ordre. L'expression science de l'évaluation indique l'accent mis

> sur le principe directeur de l'enquête systématique (Guiding principles for evaluators, 2004) et sur les normes exigeantes d'exactitude ((Joint Committee on Standards for Educational Evaluation, 1994), p.11, mis en exergue dans l'original).

Le livre de Donaldson fournit des conseils importants pour la pratique de la science de l'évaluation, en mettant particulièrement en avant la théorie du programme comme élément central de cette pratique. Il a commencé à utiliser délibérément la terminologie « science de l'évaluation » lorsqu'il a constaté que l'évaluation était traitée comme une science de second ordre dans les milieux universitaires. Depuis qu'il se réfère à ce qu'il fait comme science de l'évaluation, il rapporte avoir gagné en crédibilité auprès des scientifiques avec lesquels il interagit dans les sciences sociales et comportementales (Donaldson, 2017). Le problème ici est autant notre identité que notre pratique.

Implications, préoccupations et mises en garde

Le scientisme et la « méthode scientifique »

La recherche scientifique ne se limite pas à une définition étroite de la méthode scientifique. De même, la science de l'évaluation n'est pas réductible à, définie par, ou limitée à certaines méthodes privilégiées. […] Il est courant de juger si l'évaluation peut être considérée comme une science en mettant en avant un critère de rigueur, mais à partir d'une définition restrictive de la rigueur comme synonyme de l'usage de méthodes expérimentales traitant de façon scientiste les questions de la causalité, de la réplication et de la généralisation. Pourtant, une littérature considérable, s'appuyant sur les travaux fondateurs de Cronbach, démontre qu'une science rigoureuse implique bien plus que des méthodes expérimentales et que l'évaluation peut et doit employer

une variété de méthodes adaptées à la situation. Dans un article de *l'Encyclopedia of Evaluation* [Encyclopédie de l'évaluation] (Mathison, 2005) perspicace sur les débats méthodologiques en matière d'évaluation, les éminents praticiens et méthodologistes de l'évaluation Jennifer Greene et Gary Henry (2005) ont conclu :

> Si les normes relatives à ce qui constitue une preuve légitime pertinente pour les décisions politiques et la poursuite des programmes sont trop restrictives, excluant toutes autres preuves que celles issues d'expérimentations aléatoires à grande échelle, nous n'obtiendrons que des informations très limitées sur un très petit nombre de programmes.... Un exemple typique est la pression actuelle du gouvernement fédéral en faveur des « preuves scientifiquement fondées », qui favorise particulièrement les preuves expérimentales. Cela conduira à laisser de côté l'analyse des contextes, les récits de programmes, les commentaires de participants tirés d'études de cas, d'entretiens approfondis, de sondages ou d'autres méthodes. Cela aura aussi pour effet de faire taire les voix des nombreuses parties prenantes du programme qui peuvent et doivent être entendues. Nous, évaluatrices et évaluateurs quantitatifs et qualitatifs, devrions nous unir pour éviter que l'absence de preuves répondant à certaines normes étroites ne devienne un permis d'adopter des actions fondées uniquement sur l'idéologie ou sur une rhétorique sans borne (p.350; pour une meilleure lisibilité, les deux premières phrases ont été inversées dans cet extrait).

Des critères scientifiques étroits pour autoriser des actions entièrement fondées sur l'idéologie ou la force d'une rhétorique sans borne? Ils n'auraient guère pu être plus clairvoyants. L'argument antiscience est essentiellement scientifique : il n'y a pas de preuve scientifique absolue et définitive que le tabagisme provoque le cancer, ou que les vaccinations ne provoquent pas l'autisme, ou que les humains sont à l'origine du

changement climatique, ou que le racisme est associé aux meurtres de la police, ou…, ou…, ou… Comme antidote au scientisme étroit, Greene et Henry (2005) appellent la communauté des évaluatrices et évaluateurs à :

> s'unir dans notre engagement à mettre en œuvre et faire connaître notre acceptation durement acquise de multiples méthodes et de multiples façons de savoir, à revendiquer la contribution des sciences sociales aux politiques publiques et aux programmes sociaux et à recentrer le débat sur les valeurs plutôt que sur la méthode. Nous pourrons ainsi réorienter notre expertise et notre énergie collectives en évaluation au service de l'amélioration sociale démocratique et de la justice sociale. (p.350)

Ainsi, je tiens à préciser que l'affirmation du statut scientifique de l'évaluation n'a pas pour but de préconiser une conception étroite de la rigueur méthodologique. Comme le montrent les écrits sur la science de l'évaluation de Donaldson (2007) et de Pawson (2013), sans parler de Patton (Patton 2008, 2012, 2015), la science de l'évaluation est éclectique, pluraliste et mixte sur le plan des méthodes. La rigueur réside dans le raisonnement plus que dans les méthodes. […]

Des pratiques non scientifiques de l'évaluation

Certaines activités liées à l'évaluation – comme le monitoring de routine, le feedback axé sur l'amélioration des apprentissages, les checklists et les rapports d'activité, ainsi que les évaluations exclusives non publiées – peuvent ne pas répondre aux critères de la recherche scientifique. Ces pratiques sont toutefois des applications de la science de l'évaluation. Elles se fondent sur les méthodes et les connaissances de l'évaluation, et les appliquent. Une perspective parallèle met l'accent sur la nature technologique de la médecine. Dans son ouvrage *History of Medicine*

[Histoire de la médecine], Jacalyn Duffin (2010) affirme que « la médecine n'est pas une science mais plutôt une technologie appliquée, ou un art, qui fait largement usage de la science » (p.65).

La production de connaissances en matière d'évaluation et l'élaboration de théories et de méthodes peuvent donc être considérées comme une science de l'évaluation, tandis que l'application des connaissances, des théories et des méthodes d'évaluation peut être considérée comme une technologie de l'évaluation. Les deux désignations positionnent l'évaluation comme plus qu'une fonction administrative, de gestion, de conformité et de reddition de comptes. Les évaluations qui sont conçues de façon mécanique et mises en œuvre sans discernement pour répondre à un mandat de conformité ne sont ni de la science ni de la technologie. C'est ce que Peter Drucker (1959) a qualifié de « travailleurs du savoir » qui produisent des rapports plutôt que des produits manufacturés.

Éviter l'élitisme scientifique

[...] Certains considèrent les scientifiques comme élitistes, arrogants et distants. Bien sûr, certains considèrent les évaluateurs de la même manière. Pratiquer la science de l'évaluation peut ainsi apparaître comme une double dose d'élitisme. Nous nous sommes efforcés de rendre l'évaluation compréhensible, pratique, accessible et utile. Nous nous efforçons également de rendre l'évaluation diverse et inclusive. Les enseignements que nous avons tirés de ces efforts peuvent nous aider à communiquer au sujet de la science de l'évaluation et de sa pratique, notamment « en cherchant à faire entendre les voix inaudibles de la science » (Olmstead, 2017). Stewart Donaldson a expliqué dans un webinaire sur la science de l'évaluation que, s'il estime que le positionnement de l'évaluation comme science renforce la crédibilité dans le milieu universitaire, il évite cette étiquette lorsqu'il travaille avec des non-universitaires dans les écoles et les communautés.

Une science de l'évaluation par les citoyen-ne-s

[…] Une façon d'éviter l'élitisme scientifique est de promouvoir une science de l'évaluation citoyenne. J'ai récemment demandé à un groupe que j'animais s'ils préféraient être connus comme des intervenants en évaluation ou des scientifiques de l'évaluation citoyenne. La question a suscité une discussion animée et s'est conclue par une volonté d'essayer le nouveau nom. Les approches collaboratives et participatives en évaluation peuvent devenir des approches collaboratives et participatives en science de l'évaluation. […]

Une science de l'évaluation attentive aux valeurs

« Science sans conscience n'est que ruine de l'âme » (Montaigne, *Les Essais*, 1580-88).

On s'est inquiété du fait que la science, et donc par incidence, la science de l'évaluation, exclue ou marginalise les préoccupations en matière de justice sociale. La Marche pour la Science a ainsi été critiquée comme étant raciste et sexiste.

> Au cours des trois derniers mois, la communauté scientifique, qui est en grande partie blanche, hétérosexuelle, cisgenre, valide et masculine, a débattu avec acharnement de la nature politique de la Marche face au régime de Trump, laissant les scientifiques de milieux marginalisés se sentir… encore plus marginalisés. En réponse, les scientifiques qui s'identifient comme femmes, handicapés, queer, trans, personnes de couleur, etc. ont convergé autour du hashtag #MarginSci pour prendre à parti leurs collègues racistes et sexistes. (Ama Mantey, 2017)

Cette critique fait qu'il est essentiel d'examiner si faire de la justice sociale une priorité scientifique et un objet d'évaluation scientifique pourrait accroître l'attention et la compréhension du rôle du savoir et de la science dans la promotion de la justice sociale et dans la lutte contre le racisme et le sexisme. Pour un argument scientifique à l'appui de la justice sociale, voir le *Qualitative Manifesto* (Denzin, 2010). Bien qu'il soit axé sur l'enquête qualitative, son argument fondamental en faveur de la science à l'appui du changement sociétal et de la justice sociale est généralement applicable.

[…] La science de l'évaluation peut et doit intégrer une dimension morale. Parler de science de l'évaluation ne doit pas conduire à renforcer une approche technoscientifique étroite. La science de l'évaluation doit se préoccuper à la fois de bien faire les choses et de faire des choses bonnes. L'appel de Tom Schwandt (2004) à un discours moral dans l'évaluation et l'appel de Scriven à une infusion éthique comme troisième révolution de l'évaluation prennent une importance particulière dans un monde où les forces idéologiques antiscientifiques menacent de saper à la fois la moralité et l'éthique.

Science de l'évaluation, vérité et qualité

La mission souvent revendiquée de l'évaluation à dire la vérité au pouvoir prend une acuité particulière à l'ère de la post-vérité, et s'élargit pour inclure le fait de dire la vérité au grand public. La crédibilité de la science pour rechercher et dire la vérité dépend du contrôle de la qualité. Ni la science, ni les scientifiques, ne sont intrinsèquement bons. La « mauvaise science » (Goldacre 2009) n'est que trop courante et doit être exposée et corrigée pour maintenir l'intégrité scientifique. […]

Conclusion

Si nous voulons que l'évaluation soit reconnue comme une branche de la science parmi d'autres, nous devons commencer par reconnaître notre fondement scientifique. Si nous voulons obtenir un soutien pour financer l'évaluation, nous ferions bien de faire cause commune pour le financement de toutes les branches de la science, qu'elles soient fondamentales ou appliquées. Si nous devons jouer notre rôle dans la lutte contre les attitudes et les actions antiscientifiques, qui sont par nature aussi des attitudes et des actions anti-évaluation, alors nous ferions bien de faire cause commune avec d'autres scientifiques.

Revendiquer l'évaluation comme science renforce notre crédibilité, notre responsabilité, notre capacité, notre utilité et notre efficacité, tout en communiquant notre rôle de façon plus claire et crédible à celles et ceux qui apprécient la science mais qui n'ont pas pensé à l'évaluation en tant qu'activité scientifique. Le positionnement de l'évaluation en tant que science peut aussi avoir des conséquences sur la façon dont nous sommes perçu-e-s, traité-e-s, et situé-e-s dans les établissements universitaires, les organismes gouvernementaux, ainsi que par les bailleurs de fonds et les utilisateurs et utilisatrices de l'évaluation. [...]

Bibliographie

Ama Mantey, Jane. « #MarginSci: The March for Science as a Microcosm of Liberal Racism », *The Root*, 20 avril 2017. En ligne : https://www.theroot.com/marginsci-the-march-for-science-as-a-microcosm-of-lib-1794463442

Blackburn, Simon. 2005. *Truth: A guide*. New York: Oxford University press.

Denzin, Norman K. 2010. *The Qualitative Manifesto*. London: Routledge.

Donaldson, Stewart I. 2007. *Program theory-driven evaluation science: Strategies and applications*. Mahwah: Lawrence Erlbaum.

Donaldson, Stewart I. 2017. « Evaluation science ». Présenté à AEA eStudy webinar.

Drucker, Peter F. 1959. *The landmarks of tomorrow*. New York: Harper and Row.

Duffin, Jacalyn. 2010. *History of medicine*. 2e éd. Toronto: University of Toronto Press.

Goldacre, Ben. 2009. *Bad science*. London: Fourth Estate.

Greene, Jennifer C., et Gary T. Henry. 2005. « Qualitative-quantitative debate in evaluation ». in *Encyclopedia of evaluation*, édité par S. Mathison. Thousand Oaks: Sage Publications, p. 345-50.

Joint Committee on Standards for Educational Evaluation. 1994. *The program evaluation standards: How to assess evaluations of educational programs*. 2e éd. Thousand Oaks: Sage Publications.

Mathison, Sandra, éd. 2005. *Encyclopedia of evaluation*. Thousand Oaks: Sage Publications.

Olmstead, Molly. 2017. « Seeking the unheard voices of science: How science journalists consider diversity when finding sources ».

Patton, Michael Q. 2008. *Utilization-focused evaluation*. 4e éd. Los Angeles: Sage Publications.

Patton, Michael Q. 2012. *Essentials of utilization-focused evaluation*. Los Angeles: Sage Publications.

Patton, Michael Q. 2015. *Qualitative research & evaluation methods*. 4e éd. Los Angeles: Sage Publications.

Pawson, Ray. 2013. *The science of evaluation: A realist manifesto*. London: Sage Publications.

Schwandt, Peter. 2004. *Evaluation practice reconsidered*. New York: Peter Lang.

Scriven, Michael. 1976. *Reasoning*. New York: McGraw-Hill.

Scriven, Michael. 2004. « Reflections ». in *Evaluation roots: Tracing theorists' views and influence*, édité par M. C. Alkin. Thousand Oaks: Sage Publications, p. 183-95.

Scriven, Michael. 2008. « The concept of a transdiscipline: And of evaluation as a transdiscipline ». *Journal of Multidisciplinary Evaluation*, 5(10) : 65-66.

Scriven, Michael. 2013. « Conceptual resolutions and evaluation: Past, present, and future ». in *Evaluation roots*, édité par M. C. Alkin. Los Angeles: Sage Publications, p. 167-79.

Shadish, William R., Thomas D. Cook et Laura C. Leviton. 1991. *Foundations of program evaluation: Theories of practice*. Newberry Park: Sage Publications.

Le regard d'Yves Gingras

YVES GINGRAS

À titre de directeur scientifique de l'Observatoire des Sciences et des Technologies (OST) au Canada depuis vingt-cinq ans, j'ai participé – avec mon équipe – à de nombreuses évaluations portant sur des programmes ou des organisations liés à des activités de recherche scientifique, technologique ou d'innovation. Je ne me suis jamais demandé si l'évaluation était une « science », car la grande polysémie de ce terme rend à mon avis une telle discussion vaine et surtout stérile, en ce sens qu'elle ne génère pas de gain réel dans la pratique, même si elle peut avoir une valeur au plan rhétorique en contexte universitaire ou pour se donner de la crédibilité auprès des commanditaires.

Dans leur introduction aux textes, les autrices et auteurs invoquent le Grand Robert de la langue française qui définit la science comme un « ensemble de connaissances » obtenues par une méthode déterminée sur un objet donné. La définition proposée fait aussi référence au caractère universel, objectif et vérifiable des résultats. Une telle définition reste cependant vague, car toute « méthode » n'est pas nécessairement valide. Les parapsychologues, par exemple, disent détenir une « méthode » pour détecter les fantômes, mais peu la trouvent crédible (Collins et Pinch, 1979). Mieux vaut donc éviter le piège de ce que j'appelle le substantialisme linguistique qui porte à chercher l'essence d'un mot par le seul jeu du verbe être : demander si l'évaluation *est* une science, présuppose en effet qu'il existe un être préexistant ayant une essence immuable. En matière de langage, je suis plutôt partisan du nominalisme qui considère les définitions comme arbitraires et qu'il faut simplement s'assurer qu'on fait le même usage d'un mot donné lors d'une discussion. Ainsi, quand on dit qu'une personne est « un puit de science » le mot renvoie à une somme de connaissances. Mais une « connaissance » n'est pas une « science », tout comme un « savoir » – mot usuel en français

depuis Michel Foucault mais plus vague que « connaissance », traduction habituelle de « knowledge » – n'est pas nécessairement « scientifique ». Quand on dit que tel résultat est « scientifique » on renvoie plutôt au fait qu'il a été obtenu de manière « rigoureuse » et donc logique en suivant une méthode conforme à la nature de cet objet.

À la définition du Grand Robert, je préfère donc la suivante, à la fois plus simple et plus englobante : une science vise à « rendre raison des phénomènes par des causes naturelles » (Gingras, 2017). Pour rendre raison et donc expliquer un phénomène, on doit utiliser des concepts spécifiques à un type d'objet. Ces concepts diffèrent selon qu'on veuille rendre raison du comportement des atomes, des humains, des fourmis, des galaxies, etc. De même, les méthodes d'observation ou d'expérimentation varieront selon les objets mais le but reste le même : comprendre et expliquer les propriétés de ces objets. C'est en combinant les concepts avec les données empiriques et les lois établies, qu'on peut, éventuellement, en arriver à formuler des théories permettant de *rendre raison* d'un certain ordre de phénomène : attraction électrique, psychoses, émeutes, suicides, décisions politiques, etc.

À la lumière de cette conception de la science, on devine que l'évaluation n'est pas une science. Qu'est-ce alors que l'évaluation? Il faut je crois distinguer ici trois niveaux trop souvent amalgamés dans l'expression « recherche évaluative ». D'abord, « faire de l'évaluation », est une pratique professionnelle définie par une fin, à savoir mesurer les effets produits par une action : un programme, un projet, une politique, etc. Implicitement ou explicitement l'évaluation se fait en vue d'une action à entreprendre à la lumière des résultats obtenus. En principe, on n'évalue jamais pour rien, même si les résultats ne sont pas toujours utilisés – surtout quand ils déplaisent au commanditaire! D'autre part, on peut faire de la recherche pour concevoir, améliorer et adapter des méthodes utilisées en évaluation, sans faire soi-même des évaluations au sens défini précédemment. On est alors plutôt méthodologue. Je me suis, par exemple, intéressé aux indicateurs souvent utilisés en évaluation pour démontrer que plusieurs sont inadéquats à leur objet et ne devraient

donc pas être utilisés dans des évaluations (Gingras, 2017). Enfin, on peut faire de la recherche *sur* l'évaluation et en faire l'histoire, la sociologie ou se demander, du point de la psychologie des organisations par exemple, quels sont les effets des évaluations sur les agents qui en sont l'objet. On prendra alors l'évaluation comme objet d'analyse du point de vue d'une discipline classique visant à rendre raison d'un phénomène (ici l'évaluation) et de ses effets psychologiques, organisationnels, sociaux ou politiques.

L'évaluation existe donc d'abord en fonction d'une fin pratique qui vise à éclairer une prise de décision rationnelle concernant un programme ou une politique spécifique et localisée dans l'espace dont la mise en place visait à produire certains effets qu'il s'agit alors de mesurer pour savoir s'ils se sont bien réalisés et si oui à quel degré. Ses objets d'analyse étant potentiellement infinis et relevant en principe de tous les secteurs de la société, on comprend que l'évaluation n'a pas véritablement d'objet spécifique à proprement parler, sauf à se placer à un très haut degré d'abstraction pour définir des classes d'objets associés à des classes de méthodes d'évaluation. Mais un tel degré de généralité, déconnecté des objets concrets évalués dans leur spécificité me semble aussi peu utile qu'une « théorie générale des systèmes » telle que von Bertalanffy pouvait, par exemple, l'envisager au milieu des années 1960.

L'évaluation répondant à une demande ciblée et d'origine externe (oublions ici les autoévaluations), il me semble donc non seulement irréaliste mais absurde de penser avec Michael Scriven que l'évaluateur peut dicter ou imposer ses propres valeurs et ainsi « rejeter une conception de l'évaluation comme aide à la décision ». Ou alors il n'évaluera que les projets dont il approuve les fins. Mais la plupart des professionnels n'ont probablement pas ce luxe. On peut même penser que c'est là confondre les niveaux d'action : d'une part, faire la promotion de certaines valeurs ou finalités sociales et, d'autre part, mesurer les effets réels d'un programme auquel le commanditaire assignait des objectifs dont il veut maintenant mesurer s'ils ont été atteints ou non. À mon avis, évaluer un programme est une opération analytiquement distincte

de celle visant à le mettre au point. Évaluer consiste en effet à tenter de répondre aux questions que posent les commanditaires. Pour cela, la personne appelée à faire une évaluation fera comprendre aux commanditaires qu'on ne peut pas évaluer sans connaître la finalité qui était visée par la mise en place du programme ou de la politique. Elle lui expliquera également, si nécessaire, que certains indicateurs, pourtant suggérés, ne sont pas valides pour mesurer l'effet d'une action donnée. Idéalement, elle contribuera ainsi à ce que les responsables des programmes les construisent de manière aussi rationnelle que possible en identifiant bien au départ l'objectif visé par cette action. Bien que les personnes chargées de l'évaluation puissent être consultées en amont sur les variables et indicateurs potentiels mesurables du programme mis en place, il est évident que ce ne sont pas elles qui fixent les objectifs du programme, lesquels peuvent être infiniment variés selon les secteurs visés de la société. Par contre, au plan éthique, les personnes responsables de l'évaluation ne peuvent accepter d'évaluer à partir de critères et d'indicateurs qu'elles savent être inadéquats dans le contexte du programme et des objectifs qui étaient visés.

L'évaluation devant mesurer les effets réels d'une grande variété d'actions dans divers secteurs de la société, il va de soi qu'elle emprunte ses méthodes à plusieurs disciplines. Mais, contrairement à ce que suggère Scriven, cela n'en fait pas une « transdiscipline », surtout que ce terme est mal défini et que l'exemple des statistiques est très mal choisi. En effet, les statistiques sont une spécialité bien institutionnalisée relevant de la discipline des mathématiques et qui s'enseigne le plus souvent dans un département de mathématiques (Gingras, 2020 : 48-52). Le fait qu'elle ait des applications multiples dans d'autres disciplines ne change pas son statut car il ne faut pas confondre l'objet de recherche et ses domaines d'application. Après tout, l'arithmétique est aussi une spécialité mathématique utilisée dans tous les domaines et cela n'en fait pas une « transdiscipline ».

En somme, quand on invoque l'idée de « science » on veut généralement dire que les évaluations se font de manière rigoureuse, logique, en utilisant les outils de mesure appropriés à l'objet. Cette profession peut détenir certains concepts en propre et avoir son jargon spécifique mais sa finalité n'est pas de rendre raison d'un type de phénomène mais de mesurer les effets d'une action. Sa force réside dans sa capacité à identifier les bonnes méthodes pour mesurer les effets réels du programme étudié. Mais comme les effets sont le plus souvent multiples et imprévisibles, l'évaluation doit aussi choisir quoi mesurer en priorité et donc identifier les objectifs prioritaires qui étaient visés par la mise en place du programme. L'évaluation peut bien sûr contribuer à faire apparaître des effets non prévus mais tout de même désirables. Chose certaine, tous les effets possibles ne peuvent pas être mesurés.

Au plan sociologique, l'évaluation est clairement un domaine de pratique professionnel, avec ses sociétés savantes et ses revues. Comme le rappelle le texte de Gary Cox, le domaine a été fortement stimulé par les gouvernements qui, dans la plupart des pays, ont imposé l'obligation d'évaluer les différents programmes qu'ils mettent en place. C'est cette demande externe qui impose son caractère pratique à l'évaluation. Par contre, son insertion dans le champ universitaire génère nécessairement un discours de « scientifisation » pour justifier sa présence aux côtés de disciplines bien ancrées et se définissant comme des sciences à portée générale et explicative et relativement autonome par rapport aux demandes externes. Alors que dans les disciplines les plus autonomes, l'audience est surtout composée de personnes elles-mêmes productrices de connaissances, les rapports d'évaluation répondent à une demande externe et s'adressent au client ou à la cliente et non pas d'abord aux autres personnes pratiquant des évaluations. L'idéologie professionnelle inhérente à l'université moderne humboldtienne postule une relation nécessaire entre enseignement et recherche. Cette exigence structurelle entraine toute pratique professionnelle qui s'y intègre à développer un discours sur l'importance de la recherche académique (Gingras, 1991). Si l'enseignement de l'évaluation va de soi – car il y a bien des connaissances

à acquérir en cette matière –, la recherche n'a vraiment d'importance symbolique que pour la fraction universitaire des experts en évaluation. Les produits de l'évaluation étant très diversifiés dans leurs objets, localisés dans des contextes géographiques précis et ayant une visée pratique d'aide à la décision, ils n'ont pas tous vocation à finir en article de revue savante. La tendance à tenter de monter en généralité pour produire des « théories » à partir de cas spécifiques est ainsi un effet structurel d'appartenance à un monde académique dans lequel la hiérarchie des disciplines place la théorie en haut de l'échelle et la pratique et les « études de cas » non suffisamment théorisées en bas de l'échelle.

Comme le note Michael Quin Patton dans sa contribution, la principale raison d'invoquer le statut scientifique de l'évaluation est de « surmonter la réputation négative de la profession dans certains contextes ». C'est pour les mêmes raisons de prestige social que l'on parle des « Sciences de l'administration », alors qu'en pratique ces savoirs empruntent beaucoup à la sociologie et la psychologie. De plus, la connotation du mot « science » utilisé par Patton est ici essentiellement celle de rigueur méthodologique, empirique et logique associée à la mesure d'un phénomène et non pas celle évoquée dans le Grand Robert de la langue française. Il est significatif à cet égard que lorsqu'il cherche à caractériser ce qu'il nomme au singulier « la science de l'évaluation », les mots qui suivent sont « méthodes scientifiques rigoureuses » et « systématiques ». Mais on peut se demander si ce n'est pas là un pléonasme car « science » connote déjà l'idée de rigueur et de méthode systématique.

En fin de compte, la question centrale est, à mon avis, moins de déterminer si l'évaluation est une « science » dans au moins un des nombreux sens de ce mot socialement prisé – la réponse étant évidemment positive – que de savoir si elle a les moyens de démontrer à ses utilisateurs et utilisatrices que les résultats des évaluations conduites selon les règles de l'art sont robustes et donc résistent aux critiques méthodologiques ou conceptuelles qu'on peut leur adresser. Et l'on peut faire cela en appliquant essentiellement les « principes des méthodes

de recherche scientifique », que John Stuart Mill exposait déjà dans son *Système de logique déductive et inductive* de 1843. Car, fondamentalement, et derrière la plus grande sophistication des diverses techniques d'analyse et de mesure élaborées depuis, c'est bien encore par les méthodes des concordances, des différences et des variations concomitantes que l'on peut établir empiriquement une conclusion sur les causes et les effets, et distinguer ainsi clairement ce qui est vraiment le cas de ce que l'on voudrait qui soit le cas.

La démarche d'évaluation peut finalement être scientifique en ce sens méthodique et mener à des conclusions valides, sans que l'on doive se tourmenter plus avant pour savoir si elle a tous les attributs d'une véritable « science » qui mérite d'exister dans l'enceinte universitaire.

Bibliographie

Collins, Harry M. et Trevor J. Pinch. 1979. « The construction of the paranormal: Nothing unscientific is happening ». *The Sociological Review* 27(1) : 237-70. doi : https://doi.org/10.1111%2Fj.1467-954X.1979.tb00064.x.

Gingras, Yves. 1991. « L'institutionnalisation de la recherche en milieu universitaire et ses effets ». *Sociologie et sociétés* 23(1) : 41-54. doi : https://doi.org/10.7202/001297ar.

Gingras, Yves. 2014. *Les dérives de l'évaluation. Du bon usage de la bibliométrie*. Paris : Raisons d'agir.

Gingras, Yves. 2017. « Qu'est-ce qu'une science? » in *Qu'est-ce que la science… Pour vous*, *Sciences & philosophie*, édité par M. Silberstein. Paris : Éditions matériologiques, p. 119-24.

Gingras, Yves. 2020. *Sociologie des sciences*. 3e éd. Paris : Presses universitaires de France.

V. LA DIVERSITÉ DES APPROCHES PARADIGMATIQUES

Introduction : la pluralité des approches paradigmatiques

VALÉRY RIDDE, THOMAS DELAHAIS, AGATHE DEVAUX-SPATARAKIS ET ANNE REVILLARD

À l'instar de la santé publique ou de la médecine par exemple, l'évaluation ne relève pas seulement de la recherche (partie Sciences), mais aussi et surtout de la pratique, de la réalisation d'activités qui visent à porter un jugement (partie Valeurs) ou à comprendre une intervention ou une politique, quel que soit le domaine concerné (éducation, justice, politique, transport, etc.). Or, comme nous l'a appris Pierre Bourdieu, un tel champ est nécessairement un lieu de conflits et d'enjeux de pouvoir (Bourdieu, 2001). La pratique de l'évaluation n'en fait pas l'économie, notamment lorsqu'il s'agit d'appréhender des questions complexes comme celle de la causalité, laquelle se trouve au cœur de l'analyse de l'efficacité des interventions. En effet, comment affirmer devant tel-le responsable politique que son projet pilote visant à fournir un revenu minimal à ses concitoyennes et concitoyens, leur a permis de sortir de la pauvreté et de trouver un emploi, et de ce fait rend essentiel son déploiement à une très large échelle? Comment étayer cette démonstration, et apporter la preuve que le revenu minimal a permis un retour à l'emploi? Nous sommes là au centre de la compréhension de la nature de la preuve, de nos schémas de pensée et partant, de nos manières de concevoir le monde.

Les paradigmes

Notre façon d'appréhender le monde est souvent résumée par un terme relativement galvaudé aujourd'hui, car utilisé trop souvent en dehors de son sens premier, celui de paradigme. Il existe de multiples acceptions de ce concept, lequel est utilisé par exemple dans le domaine de l'étude des politiques publiques pour évoquer des ruptures majeures (Hall, 1993), en une reprise des propositions du philosophe des sciences Thomas Kuhn au sujet des révolutions scientifiques (Kuhn, 2016). La notion de paradigme renvoie essentiellement à quatre dimensions interreliées : l'épistémologie (relation que l'on entretient avec les données), la méthode (outils d'évaluation), l'ontologie (nature du monde et manipulation possible des objets de l'évaluation), la téléologie (finalité des évaluations). Nous avons évoqué en détail ces dimensions dans un ouvrage francophone concernant l'évaluation (Dagenais et Ridde, 2012).

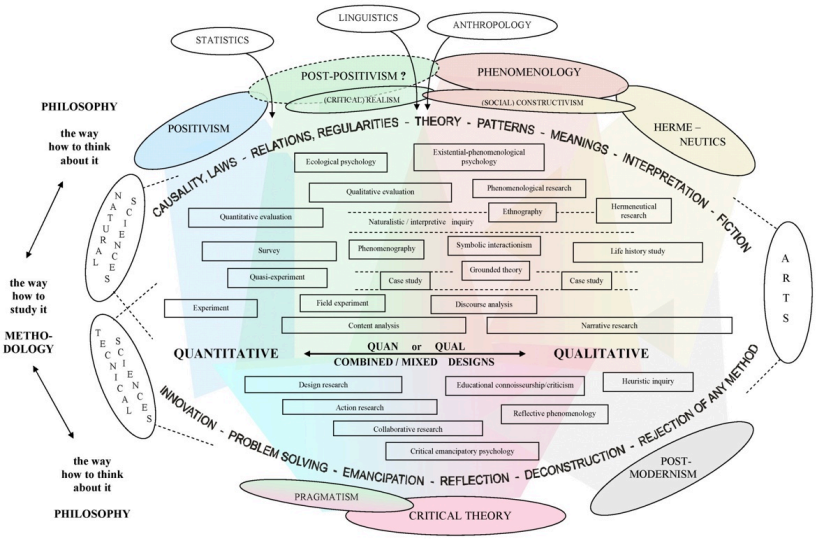
Si l'on se réfère au développement historique de l'évaluation, les paradigmes concernés sont pour l'essentiel les paradigmes scientifiques. Sans vouloir simplifier exagérément, il est possible de les résumer en trois catégories : le positivisme, le réalisme et le relativisme. Ces catégories, comprises comme des idéaux types n'épuisant pas la complexité, font appel à des représentations du monde différentes et donc, si l'on reprend l'une des quatre dimensions précédentes, à l'utilisation de méthodes particulières. Mais il ne faut pas s'y méprendre, une évaluation est d'abord organisée pour répondre à des questions concernant une intervention et non pas – ce qui est essentiellement le propos des équipes de recherche – pour mobiliser des méthodes. Il n'existe pas plus de méthode parfaite que d'approche paradigmatique idéale. L'important, dans le domaine de l'évaluation, est de trouver les moyens techniques et intellectuels de disposer des données nécessaires pour répondre aux questions d'évaluation, et pour ainsi étayer les décisions qui seront prises.

Les textes rassemblés dans ce chapitre permettent de mettre en évidence des différences, des subtilités, mais aussi des ressemblances, et de montrer l'évolution des débats sur ces questions. Si l'on s'attarde aux deux extrêmes du *continuum*, les personnes ancrées dans une approche positiviste auront tendance à choisir une démarche méthodologique préalablement à l'évaluation (en utilisant surtout des données quantitatives), sans la faire évoluer par peur de biaiser ou d'influer sur les résultats, eu égard à leur croyance en l'objectivité. À l'autre extrême, les relativistes vont principalement prendre en compte les données qualitatives, et adapter en permanence leurs outils afin de mieux s'adapter au contexte et aux données – lesquels sont en transformation constante-, abandonnant le rêve de l'objectivité pour s'orienter vers une objectivation assumée de ce qu'ils observent.

D'un capharnaüm de paradigmes au pragmatisme

Mais la vie, et surtout la pratique de l'évaluation ne se résument pas à cette dichotomie, surtout si l'on convient qu'il s'agit avant tout de répondre à des questions spécifiques concernant une intervention. Ce sont les questions qui doivent guider les évaluations, non pas les méthodes. Il existe un réel continuum, sinon un capharnaüm (cf. figure ci-dessous), dans les approches paradigmatiques, remises notamment au goût du jour par les tenants du pragmatisme et des méthodes mixtes. Katrin Niglas (2010) a tenté une représentation visuelle de cette complexité impressionnante, qui montre l'imbrication des paradigmes, des méthodes et des disciplines sans pour autant en cloisonner les approches.

Figure 1 : Un modèle multidimensionnel des méthodes et des paradigmes de recherche



Source : Katrin Niglas http://www.tlu.ee/~katrin/mmrm/skeem_handbook.jpg

Reprenant l'appel des Français Daniel Schwartz et Joseph Lellouch (Schwartz et Lellouch, 2009) de la fin des années 1960 à plus de pragmatisme, les tenants de cette perspective dans le domaine de l'évaluation réclament moins de dogmatisme. S'attacher à un paradigme ou à des méthodes en particulier, permet rarement de répondre aux objectifs des évaluations. Cependant, il ne faut pas sous-estimer le poids de ces enjeux dans les débats, lesquels ont fait rage entre les tenants de l'évaluation qualitative ou quantitative dans les années 1980 et 1990. Puis, il semble qu'ils se soient estompés au tournant du siècle, au profit de la vision commune d'un primat des objectifs de l'évaluation sur le choix des méthodes (Mark, 2003). C'était pour mieux renaître, quelques années plus tard, avec l'offensive de partisans et partisanes des méthodes

expérimentales par assignation aléatoire (*randomistas*) tendant à nier la scientificité de toute autre méthode (Donaldson, 2009). Si le bilan du recours à ces méthodes dans les années 2000 n'apparaît pas bien différent de celui établi dans les années 1980 (Heckman, 2020), il aura fallu plus de 10 ans pour que la communauté évaluative fasse apparaître de nouvelles approches causales et les méthodes pour les rendre opérationnelles (Stern *et al.*, 2012).

Les controverses restent donc nombreuses, mais les évolutions sont possibles et intéressantes, comme en témoignent les textes rassemblés dans ce chapitre.

Fondements : de quelle causalité parle-t-on?

Les débats les plus anciens et les plus virulents sur les questions paradigmatiques dans le domaine de l'évaluation concernent surtout la question de l'efficacité des interventions et la nature de la preuve attestant que les objectifs des interventions sont atteints (la causalité). Ces débats ont aussi concerné d'autres sujets, comme celui du rôle de l'évaluateur ou évaluatrice, ou de la manière dont les personnes concernées par l'intervention participent à l'évaluation (partie Evaluatrice). Mais les discussions les plus engagées ont de tout temps, encore aujourd'hui, porté sur la manière dont il est possible d'affirmer qu'une intervention a été efficace. Or, pour ce faire, il faut se mettre d'accord sur la manière dont on comprend la causalité, autrement dit sur le fait que l'intervention Y puisse être tenue responsable des changements X, et sur les processus et les données à mobiliser pour rendre compte des effets X, à la suite de l'organisation de l'activité Y. À l'instar des paradigmes, il existe une myriade de points de vue sur ce sujet.

Pour simplifier, jusqu'à la fin des années 1990, le débat oppose deux camps. D'une part, celui des tenantes et tenants de la causalité par le repérage d'une succession d'événements montrant que l'effet attendu

d'une intervention est produit par des actions spécifiques, précises et préalables à cette dernière. La personne qui évalue va alors chercher à contrôler ces processus et ces variables contextuelles, pour estimer les effets de cette relation relativement linéaire entre causes et conséquences. Les approches expérimentales discutées par Donald Campbell et Julian Stanley (**texte 1**) sont au cœur de cette perspective. Leur texte offre une riche présentation de leurs origines et fondements ainsi que du développement subséquent des approches plus pragmatiques, mais toujours ancrées dans cette même vision du monde, et qualifiées de quasi expérimentales.

D'autre part, et selon une perspective très différente, nous trouvons les tenant-e-s d'une causalité qui est en mesure de prendre en compte les éléments de contexte et d'environnement, pour qui la vision linéaire et immuable d'une relation entre une intervention et ses effets n'existe pas dans le monde réel. Cette vision dite générative (ou explicative) de la causalité postule que les effets des interventions s'expliquent par la volonté et le raisonnement des personnes qui les produisent. Ces débats ont récemment resurgi dans le monde de l'évaluation, mais ils sont très anciens, comme le montre le texte de Lawrence Mohr (**texte 2**) concernant la mobilisation des approches qualitatives dans l'évaluation de l'impact des interventions. Ce texte est fondamental pour comprendre la notion de causalité et les enjeux paradigmatiques de la compréhension de la comparaison entre, d'un côté, les effets provoqués par une intervention, et de l'autre, ce qui se serait produit sans cette action, le fameux concept de « contrefactuel ». Il est aussi une introduction à l'approche de « modus operandi » proposée par Michael Scriven, qui agit comme un détective à la recherche d'indices et de traces, en éliminant au fur et à mesure de l'évaluation certaines causes possibles des effets constatés de l'intervention (voir aussi la démarche du « process tracing ») (Maillet et Mayaux, 2018). Cette forme de raisonnement est évidemment différente de celle du « contrefactuel » pour juger de l'efficacité d'une intervention, ainsi que l'illustrent ces deux textes.

Controverses : les méthodes expérimentales par assignation aléatoire

« La guerre de la causalité fait toujours rage » : tel est le début de l'article que Scriven écrit en 2008 (**texte 3**). À ce sujet, les controverses, pour ne pas utiliser un vocabulaire guerrier redevenu à la mode en temps de pandémie, remontent à loin. En France en particulier, le retour de l'approche expérimentale pour l'évaluation de l'efficacité des interventions leur a permis de refaire surface, alors qu'elles sont très anciennes, sinon consubstantielles à la pratique de l'évaluation. Depuis l'expérimentation du revenu de solidarité active, jusqu'à la médiatisation d'un prix Nobel de l'économie du développement, en passant par les débats contemporains sur l'efficacité d'un médicament contre un nouveau virus, il s'agit du même type de démarche dont il est question pour évaluer l'efficacité d'une intervention : les essais contrôlés randomisés (ECR), autrement appelés en français méthodes expérimentales par assignation aléatoire. Même si les controverses sur la question paradigmatique sont multiples dans les démarches d'évaluation, nous avons choisi d'en rendre compte à propos des ECR, car en cette matière, le débat est permanent. Il ne sera certainement jamais terminé, mais il nous paraît important que la lecture de cet ouvrage permette de l'appréhender dans la durée.

Alors que les textes cités précédemment permettent de comprendre les fondements des approches expérimentales, Scriven (**texte 3**) offre une mise en perspective intéressante des ECR à partir de son analyse… expérimentée. Il présente une dizaine de propositions concernant cette démarche, tant au plan paradigmatique que méthodologique. Pour chacune, il soulève des controverses afin de susciter une discussion ouverte tout en prenant de la distance, et, en conclusion, affirme que loin d'être une querelle habituelle d'universitaires et d'égos, il s'agit là d'un débat essentiel à la pratique de l'évaluation. La compréhension de la nature de la causalité est au centre des réflexions de Scriven, et surtout,

de la manière de l'appréhender dans le contexte d'une évaluation crédible, sans tomber dans le dogmatisme des approches méthodologiques. Il ne s'inscrit pas dans une approche paradigmatique particulière et son approche de la causalité est logique et empirique. Dans un texte un peu plus ancien, Joseph Maxwell (**texte 4**) aborde la controverse concernant l'utilisation de données qualitatives pour l'évaluation de l'efficacité des interventions et partant de la mise en évidence d'une relation causale. Nous aurons compris que les tenants de la causalité « successionniste » souhaitent mesurer des variables dépendantes (les effets) et contrôler les facteurs contextuels. Dans cette perspective, ils remettent évidemment en question la capacité de se servir du discours des acteurs sociaux ou de l'observation de la réalité pour l'analyse causale, ce que réfute Maxwell. Ce dernier revient sur les différentes approches paradigmatiques de la causalité et des controverses à leur sujet, pour introduire une démarche évaluative qui va devenir de plus en plus populaire dans les années 2000: l'approche réaliste. Cette démarche initialement proposée par des sociologues et des criminologues anglais a fait l'objet de nombreuses contributions en français (Robert et Ridde, 2014), nous n'y reviendrons donc pas en détail. Mais l'intérêt du texte de Maxwell est de s'inscrire dans ces débats et de montrer qu'ils ne sont pas nouveaux. La philosophie et la sociologie des sciences sont convoquées pour comprendre la notion de causalité dans une perspective très différente de celle prônée par les tenants des ECR. Il revient aussi sur les fondements de l'approche réaliste en évaluation, comprise comme une autre solution à la dichotomie présentée en introduction de ce chapitre. Ce texte est également très utile pour appréhender les notions de processus et de théorie d'intervention comme dimensions essentielles à la compréhension de l'efficacité des interventions. En effet, la plupart du temps, les évaluations utilisant uniquement des ECR ne donnent pas d'explication quant à la manière dont les effets constatés ont été produits, contrairement à ce que propose White par exemple. C'est la fameuse notion de « boîte noire » des ECR (valable aussi parfois pour d'autres approches), source de multiples controverses depuis toujours et à propos de laquelle les textes cités dans le paragraphe suivant permettent d'entrer en détail. Mais Maxwell tente

d'approfondir la réflexion en montrant comment l'analyse des processus n'est pas cantonnée à la description de ce qui se passe dans une intervention, mais peut aussi être utile pour rendre compte d'explications causales. Pour arriver à ces explications, il nous suggère plusieurs méthodes détaillées permettant de conforter la confiance que l'on peut avoir dans les approches qualitatives pour une démarche d'inférence causale.

Perspectives : pragmatisme et réalisme?

On aura compris à la lecture des textes précédents que les débats et controverses au sujet de la manière d'évaluer l'efficacité des interventions sont anciens et nombreux. Nous avons pris l'exemple des discussions sur la causalité, car elle nous semble la plus prégnante et illustrative de la permanence de la réflexion. L'avenir de l'évaluation est certainement du côté de la mise à l'écart de tout dogmatisme disciplinaire ou paradigmatique. Pour lutter contre un virus par exemple, la biologie, l'épidémiologie ou la médecine seule ne suffisent pas. De la même façon, la compréhension d'une intervention et l'évaluation de son efficacité ne peuvent se limiter à l'usage unique d'un ECR ou d'une approche ethnographique. Il aura fallu une bonne dizaine d'années à la communauté évaluative pour s'extraire de cette opposition stérile en revenant aux fondements de la causalité. Le futur s'articule ainsi autour de quatre grandes familles causales : i) le contrefactuel et la régularité, avec le retour en force d'études basées sur des observations de grande qualité; ii) la génération, avec des approches renouvelées telles que l'évaluation réaliste; iii) la reconstitution de processus (*process tracing*) ou l'analyse de contribution; iv) la configuration, qui s'attache à repérer les conditions nécessaires ou suffisantes associées aux changements attendus (Stern *et al.*, 2012).

À l'image des personnes suggérant un recours accru aux méthodes mixtes dans le domaine de l'évaluation (Bujold *et al.*, 2018), dépassant ainsi les discussions paradigmatiques pour plus de pragmatisme, le futur est peut-être dans une voie du dépassement consistant à mêler les approches.

En effet, nous assistons depuis quelques années à un double renouveau des approches fondées sur les ECR et de l'approche réaliste, au sujet desquelles, comme nous l'avons vu plus haut, les discussions sont récurrentes. Michael Patton aimait à rappeler cette fameuse guerre des paradigmes et son pouvoir sclérosant pour la pratique de l'évaluation. Si le recours concomitant ou séquentiel à des méthodes quantitatives et qualitatives n'est pas nouveau, leur conceptualisation et l'approche réflexive à leur égard est plus récent, permettant notamment de discuter du besoin de dépassement des frontières. C'est dans ce contexte de tentative de réconciliation que les récents développements autour des ECR et de l'approche réaliste peuvent constituer une discussion intéressante sur le futur de l'évaluation.

En effet, depuis quelques années, ce double renouveau dans le champ de l'évaluation n'est pas sans poser de questions concernant les approches paradigmatiques, certains allant jusqu'à proposer, non pas d'insérer une démarche tirée du paradigme réaliste dans les ECR, mais bien de développer des ECR réalistes (**texte 5**). Comme pour les méthodes mixtes, certain-e-s pensent que les paradigmes sont irréconciliables, tandis que d'autres œuvrent pour le dialogue. Personne ne peut évidemment prédire l'avenir du champ de l'évaluation. Ce dernier sera peut-être construit par des individus qui ne souhaitent pas prendre position et s'inscriront plutôt dans un pragmatisme où les débats sur la vision du monde seront ignorés au profit d'une réponse rigoureuse aux commanditaires des évaluations (les citoyen-ne-s compris-e-s). Toujours est-il qu'il nous a semblé important de donner accès à un débat contemporain et relativement nouveau autour de la conciliation entre les ECR et l'approche réaliste en évaluation. Trois textes permettent d'instaurer un dialogue. Ils traduisent les débats sur la tentative de réconciliation entre les tenants des ECR et ceux de l'évaluation réaliste, dont on peut dire que les approches

de la causalité sont très différentes. La lecture de ces textes apporte des éléments-clefs au débat et permettra, notamment de se forger une opinion, sur la question centrale de la causalité. La pratique de l'évaluation ira-t-elle dans le sens des « puristes » de l'approche réaliste (**texte 6**) en affirmant que l'appréhension de la causalité, des mécanismes sous-jacents et des méthodes mobilisées sont incompatibles avec le paradigme dont ils se défendent? Ou alors sera-t-elle plutôt en accord avec les collègues qui proposent les ECR réalistes (**texte 5**, **texte 7**) permettant d'allier la force de la démonstration de la causalité au moyen de l'approche par assignation aléatoire des ECR, avec la compréhension de la « boîte noire » de l'intervention à l'aide de l'approche réaliste? L'une des questions centrales quant au futur de l'évaluation est donc celle – oxymorique? – de l'association entre les approches expérimentales et le réalisme critique.

Bibliographie

Bourdieu, Pierre. 2001. *Science de la science et réflexivité*. Paris : Raisons d'agir.

Bujold, Mathieu, Nha Hong, Valéry Ridde, Claude Julie Bourque, Maman Joyce Dogba, Isabelle Vedel et Pierre Pluye. 2018. *Oser les défis des méthodes mixtes en sciences sociales et sciences de la santé*. Montréal : ACFAS.

Dagenais, Christian et Valéry Ridde. 2012. *Approches et pratiques en évaluation de programmes*. Montréal : Presses de l'Université de Montréal.

Donaldson, Stewart I. 2009. « In Search of the Blueprint for an Evidence-Based Global Society ». in *What counts as credible evidence in applied research and evaluation practice*, édité par S. I. Donaldson, C. A. Christie et M. M. Mark. Los Angeles: Sage Publications, p. 2-18.

Hall, Peter A. 1993. « Policy paradigms, social learning and the state: the case of economic Policymaking in Britain ». *Comparative Politics* 25(3) : 275-296.

Heckman, James J. 2020. « Epilogue: Randomization and Social Policy Evaluation Revisited ». in *RCTs in Development: A Critical Perspective*, édité par F. Bédécarrats, I. Guérin et F. Roubaud. Oxford: Oxford University Press, p. 304-30.

Kuhn, Thomas S. 2016. *La structure des révolutions scientifiques*. Paris : Flammarion.

Maillet, Antoine et Pierre-Louis Mayaux. 2018. « Le *process tracing* : Entre narration historique et raisonnement expérimental ». *Revue française de science politique* 68(6) : 1061. doi : 10.3917/rfsp.686.1061.

Mark, Melvin M. 2003. « Toward a integrative view of the theory and practice of program and policy evaluation ». in S. I. Donaldson & M. Scriven (éds.) *Evaluating social programs and problems: Visions for the new millennium*. Mahwah, NJ: Erlbaum, p. 183-204.

Niglas, Katrin. 2010. « The multidimensional model of research methodology: An integrated set of continua ». in *Mixed methods in social and behavioral research*. Los Angeles, p. 215-36.

Robert, Émilie et Valéry Ridde. 2014. « L'approche réaliste pour l'évaluation de programmes et la revue systématique : de la théorie à la pratique ». *Mesure et évaluation en éducation* 36(3) : 79-108.

Schwartz, Daniel, et Joseph Lellouch. 2009. « Explanatory and Pragmatic Attitudes in Therapeutical Trials ». *Journal of Clinical Epidemiology* 62(5) : 499-505. doi : 10.1016/j.jclinepi.2009.01.012.

Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, et Barbara Befani. 2012. *Broadening the range of designs and methods for impact evaluations. Report of a study commissioned by the Department for International Development*. 38. Department for International Development (Dfid).

1. Protocoles expérimentaux et quasi-expérimentaux pour la recherche

DONALD T. CAMPBELL ET JULIAN C. STANLEY

[Traduit de : Campbell, Donald T. et Julian C. Stanley. 1967. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally & Company, p. 1-6. Traduction par Carine Gazier et Valéry Ridde; traduction et reproduction du texte avec l'autorisation de Houghton Mifflin.]

Dans ce chapitre, nous examinerons la validité de 16 modèles expérimentaux par rapport à 12 menaces communes qui pèsent sur la validité des inférences. Par expérience, nous faisons référence à la partie de la recherche au sein de laquelle les variables sont manipulées et leurs effets sur d'autres variables observés. Il ne s'agit pas d'un chapitre sur le protocole expérimental dans la tradition de Fisher (1925, 1935), dans lequel un expérimentateur ou une expérimentatrice ayant une maîtrise complète peut programmer des traitements et des mesures pour une efficacité statistique optimale. Dans la mesure où les protocoles de recherche examinés dans le présent chapitre deviennent complexes, c'est à cause de l'intransigeance de l'environnement et de l'impossibilité d'un contrôle complet par l'expérimentateur ou l'expérimentatrice. Bien que nous ferons ponctuellement référence à la tradition de Fisher, pour une présentation plus systématique de cette tradition nous renvoyons notamment aux livres de Brownlee (1960), Cox (1958), Edwards (1960), Ferguson (1959), Johnson (1949), Johnson et Jackson (1959), Lindquist (1953), McNemar (1962) et Winer (1962). (Voir aussi Stanley, 1957b).

Problème et contexte

McCall comme modèle

En 1923, W. A. McCall a publié un livre intitulé « Comment mener des expériences dans le domaine de l'éducation ». Le présent chapitre reprend les considérations de ce livre en les actualisant. Ainsi, il commencera par l'étudier. Dans sa préface, McCall affirme : « Il existe d'excellents livres et enseignements traitant de l'analyse statistique des données expérimentales, mais il n'y a guère d'aide sur les méthodes permettant d'obtenir des données adéquates et correctes auxquelles appliquer la procédure statistique ». Cette phrase reste suffisamment vraie aujourd'hui pour servir de fil conducteur à cette présentation. Bien que l'impact de la tradition de Fisher ait permis de remédier à la situation de façon fondamentale, son effet le plus visible semble avoir été d'élaborer une analyse statistique plutôt que d'aider à obtenir des « données adéquates et correctes ». Probablement en raison de son orientation pratique et rationnelle, et de son manque de prétention à une contribution plus fondamentale, le livre de McCall est un classique sous-estimé. À l'époque de sa parution, deux ans avant la première édition des *Statistical Methods for Research Workers* [*Méthodes statistiques pour les chercheurs*] de Fisher (1925), il n'existait aucune publication d'un niveau d'excellence comparable en agriculture ou en psychologie. Il a anticipé les méthodes orthodoxes de ces autres domaines sur plusieurs points fondamentaux. Peut-être que la contribution la plus fondamentale de Fisher a été le concept d'une mise en équivalence préexpérimentale des groupes par assignation aléatoire. Ce concept, qui revient à abandonner la stratégie (intuitivement plus attrayante, mais trompeuse) de mise en équivalence des groupes par l'appariement, a été difficile à accepter pour les chercheurs et chercheuses en éducation. En 1923, McCall avait proposé une première formulation qualitative de cette approche, en mettant en avant une première méthode d'établissement de groupes

comparables « par le hasard ». « Tout comme la représentativité peut être assurée par la méthode du hasard… l'équivalence peut être assurée par le hasard, à condition que le nombre de sujets soit suffisamment élevé » (p. 41). L'approche de Fisher avait également des précurseurs sur un autre point. Dès 1916, Thorndike, McCall et Chapman (1916) avaient introduit la conception du « carré latin », sous le terme, d'« expérience de rotation », dans les deux formes 5 X 5 et 2 X 2, c'est-à-dire une dizaine d'années avant que Fisher (1926) l'ait intégré systématiquement dans son plan expérimental, avec assignation aléatoire[1].

La façon dont McCall utilise l'« expérience de rotation » illustre bien l'importance de son livre et du présent chapitre. L'expérience de rotation est introduite non pas pour des raisons d'efficacité, mais plutôt pour obtenir un certain degré de contrôle lorsque l'assignation aléatoire à des groupes équivalents n'est pas possible. Dans le même ordre d'idées, ce chapitre examinera les imperfections de nombreux protocoles expérimentaux et plaidera néanmoins en faveur de leur utilisation dans les contextes où il n'est pas possible d'obtenir de meilleurs protocoles de recherches expérimentales. En ce sens, la majorité des protocoles de recherche discutés, y compris l'« expérience de rotation » qui n'implique pas d'assignation aléatoire, sont désignés comme des protocoles de recherches *quasi* expérimentaux.

1. Kendall et Buckland (1957) disent que le carré latin a été inventé par le mathématicien Euler en 1782. Thorndike, Chapman et McCall n'utilisent pas ce terme.

Des désillusions face à l'expérimentation dans l'éducation

Le présent chapitre est consacré à la méthode expérimentale, comme seul moyen de régler les différends relatifs aux pratiques éducatives, comme seul moyen de vérifier les améliorations de l'éducation, et comme seul moyen d'établir une tradition cumulative dans laquelle des améliorations peuvent être introduites sans risquer de faire disparaître des pratiques anciennes, mais efficaces au profit de nouveautés inférieures. Pour autant, il ne s'agit pas de prétendre que cette défense de l'expérimentation serait nouvelle. Comme l'indique clairement l'existence du livre de McCall, une vague d'enthousiasme pour l'expérimentation a dominé le domaine de l'éducation à l'époque de Thorndike, atteignant peut-être son sommet dans les années 1920. Et cet enthousiasme a fait place à l'apathie et au rejet, ainsi qu'à l'adoption de nouvelles psychologies ne se prêtant pas à la vérification expérimentale. Good et Scates (1954 : 716-21) ont ainsi rendu compte d'une vague de pessimisme, autour de 1935, même un fervent défenseur de l'expérimentation comme Monroe (1938) déclarait alors que « les contributions directes de l'expérimentation contrôlée ont été décevantes ». Plusieurs personnes bien formées dans la tradition expérimentale se sont alors tournées vers l'écriture d'essais, souvent accompagnées de la conversion du behaviorisme thorndikien à la psychologie de la Gestalt ou à la psychanalyse.

Pour éviter que cette désillusion ne se reproduise, il importe d'identifier certaines sources de la réaction précédente pour essayer d'éviter les fausses anticipations qui y ont conduit. Plusieurs aspects peuvent être notés. Tout d'abord, les affirmations concernant le niveau et le degré de progrès qui résulteraient de l'expérience étaient exagérément optimistes et s'accompagnaient d'une dépréciation injustifiée de la sagesse non expérimentale. Les partisans initiaux supposaient que les progrès dans la technologie de l'enseignement avaient été lents simplement parce que la méthode scientifique n'avait pas été appliquée. Selon cette logique,

la pratique traditionnelle était inadaptée parce qu'elle n'avait pas été produite par l'expérimentation. Lorsque les expériences se sont souvent révélées fastidieuses, équivoques, d'une reproductibilité peu fiable, ou encore lorsqu'elles aboutissaient finalement à valider les options issues de la sagesse préscientifique, et pour confirmer la sagesse préscientifique, les arguments trop optimistes sur lesquels l'expérimentation avait été justifiée ont été réduits à néant, conduisant à un rejet ou une mise à l'écart désabusée.

Cette désillusion été partagée tant par les expérimentateurs et expérimentatrices que par les participantes et participants à l'expérimentation. Les premier-e-s ont développé un conditionnement personnel à éviter les expérimentations. Pour tout chercheur ou toute chercheuse motivé-e, la non-confirmation d'une hypothèse importante est douloureuse. En tant qu'animal biologique et psychologique, l'expérimentateur ou l'expérimentatrice est soumis-e à des lois d'apprentissage qui l'amènent inévitablement à associer cette douleur aux *stimuli* et aux événements contigus. Le processus expérimental lui-même est souvent perçu comme étant le *stimulus* à l'œuvre, plus directement que la « véritable » source de frustration, c'est-à-dire la théorie inadéquate. Cela peut conduire, peut-être inconsciemment, à éviter ou à rejeter le processus expérimental. Si, comme cela semble probable, l'écologie de notre science est celle dans laquelle il existe beaucoup plus de mauvaises réponses que de bonnes, nous pouvons prévoir que la plupart des expériences soient décevantes. Nous devons en quelque sorte vacciner les jeunes expérimentateurs et expérimentatrices contre cet effet. En général, nous devons justifier l'expérimentation sur des bases plus pessimistes – non pas comme une panacée, mais plutôt comme la seule voie disponible pour un progrès cumulatif. Nous devons apprendre à nos étudiantes et étudiants à s'attendre à de l'ennui et de la déception, et à cultiver une persévérance approfondie, à l'instar des sciences biologiques et physiques. Nous devons élargir le vœu de pauvreté de nos étudiant-e-s pour y inclure non seulement la volonté d'accepter la pauvreté financière, mais aussi la pauvreté des résultats expérimentaux.

Plus précisément, nous devons étendre notre horizon temporel et reconnaître que l'expérimentation scientifique est un processus continu et multiple, plutôt qu'une démarche ponctuelle aboutissant à des résultats définitifs. Les expériences que nous réalisons aujourd'hui, si elles sont couronnées de succès, devront être reproduites et validées à d'autres moments et dans d'autres conditions avant de pouvoir devenir une partie établie de la science, pouvant être théoriquement interprétée avec confiance. En outre, même si nous considérons l'expérimentation comme langage de base de la preuve, comme seul tribunal permettant de trancher entre des théories rivales, nous ne devons pas nous attendre à ce que les « expériences cruciales » mettant en opposition des théories conflictuelles aient des résultats clairs. Lorsque l'on constate, par exemple, que les personnes compétentes qui observent la situation défendent fortement des points de vue divergents, il semble probable, *a priori*, que les deux aient observé quelque chose de valable sur la situation naturelle et que les deux représentent une partie de la vérité. Plus la controverse est forte, plus c'est probable. On peut donc s'attendre, dans de tels cas, à un résultat expérimental avec des conclusions mitigées, ou avec l'équilibre de la vérité variant subtilement d'une expérience à l'autre. La démarche plus avancée – en grande partie atteinte par la psychologie cognitive (par exemple Underwood, 1957b) – évite des expériences cruciales et étudie plutôt les relations et interactions dimensionnelles sur plusieurs degrés des variables expérimentales.

Il ne faut pas non plus négliger la diffusion, en psychologie et en éducation, de procédures statistiques considérablement améliorées. Au cours de la période de sa plus grande activité, l'expérimentation éducative s'est déroulée de façon inefficace avec des outils trop pointilleux. McCall (1923) et ses contemporain-e-s menaient leurs recherches en étudiant une seule variable à la fois. Cela s'est révélé trop restrictif par rapport au degré de complexité des processus d'apprentissage humain. Nous savons maintenant à quel point diverses contingences – dépendances à l'action

conjointe de deux variables expérimentales ou plus – peuvent être importantes. Stanley (1957a, 1960, 1961b, 1961c, 1962), Stanley et Wiley (1962), et d'autres ont souligné l'évaluation de ces interactions.

Les expériences peuvent être multivariées de deux façons. Plus d'une variable « indépendante » (sexe, niveau scolaire, méthode d'enseignement arithmétique, police et taille des caractères d'impression, etc.) peut être incorporée dans le protocole de recherche et plus d'une variable « dépendante » (nombre d'erreurs, vitesse, nombre exact, divers tests, etc.) peut être utilisée. Les procédures de Fisher sont multivariées dans le premier sens, univariées dans le second. Les statisticiens mathématiques, par exemple Roy et Gnanadesikan (1959), travaillent à des protocoles et à des analyses qui unifient les deux types de protocoles de recherches multivariés. Peut-être qu'en étant vigilant-e-s, les chercheuses et les chercheurs en éducation peuvent réduire le décalage généralement important entre l'introduction d'une procédure statistique dans la littérature technique et son utilisation dans les enquêtes de fond. Sans aucun doute, une formation plus poussée des chercheuses et chercheurs en éducation dans le domaine des statistiques expérimentales devrait contribuer à améliorer la qualité des démarches expérimentales en éducation.

Perspective d'évolution de la sagesse et de la science cumulatives

Les commentaires des paragraphes précédents, et une grande partie de ce qui suit, sont fondés sur une perspective évolutive de la connaissance (Campbell, 1959) dans laquelle la pratique appliquée et les connaissances scientifiques sont considérées comme le résultat d'un cumul de tentatives retenues sélectivement, les autres ayant été éliminées par l'expérience. Une telle perspective conduit à un respect considérable de la tradition dans la pratique de l'enseignement. Si, en effet, au fil des siècles, de

nombreuses approches différentes ont été essayées, si certaines ont mieux fonctionné que d'autres, et si celles qui ont mieux fonctionné ont donc été, dans une certaine mesure, appliquées de façon plus persistante par leurs auteurs et autrices, ou imitées par d'autres, ou enseignées aux apprenti-e-s, alors les coutumes qui ont émergé peuvent représenter un sous-ensemble précieux et éprouvé de toutes les pratiques possibles.

Mais l'aspect sélectif et avant-gardiste de ce processus d'évolution est très imprécis dans le cadre naturel. Les conditions d'observation, tant physiques que psychologiques, sont loin d'être optimales. Ce qui survit, ou est retenu, est en grande partie déterminé par le hasard pur. L'expérimentation devient à ce stade le moyen d'affiner la pertinence du processus de test, d'analyse et de sélection. L'expérimentation n'est donc pas considérée elle-même comme une source d'idées nécessairement contradictoire avec la sagesse traditionnelle. Il s'agit plutôt d'un processus d'affinage qui s'ajoute aux acquis probablement importants de la sagesse pratique cumulée. La promotion d'une science expérimentale de l'éducation n'est donc pas incompatible avec la sagesse traditionnelle.

Certains lecteurs pensent peut-être que l'analogie avec le schéma évolutif de Darwin devient compliquée avec des facteurs spécifiquement humains. Un directeur ou une directrice d'école lambda, lorsqu'il ou elle est confronté-e à la nécessité de décider d'adopter un manuel révisé ou de conserver la version non révisée plus longtemps, effectue probablement son choix sur la base de connaissances limitées. En plus de l'efficacité de l'enseignement et de l'apprentissage, de nombreuses considérations lui viennent à l'esprit. Il ou elle peut avoir raison de deux façons : garder l'ancien livre quand il est aussi bon ou meilleur que le livre révisé, ou adopter le livre révisé quand il est supérieur à l'édition non révisée. De même, il ou elle peut se tromper de deux façons : garder l'ancien livre quand le nouveau est meilleur, ou adopter le nouveau livre quand il ne vaut pas mieux que l'ancien.

Des « coûts » de plusieurs types peuvent être estimés approximativement pour chacun des deux choix erronés : (1) les coûts financiers et les coûts des dépenses d'énergie; (2) le coût, pour le directeur ou la directrice, des plaintes des enseignant-e-s, des parents et du conseil d'administration; (3) le coût pour les enseignant-e-s, les élèves et la société en raison d'une instruction de moins bonne qualité. Ces coûts en termes d'argent, d'énergie, de confusion, d'appauvrissement des apprentissages et de risque personnel doivent être comparés avec la probabilité que chacun se produise, ainsi qu'avec la probabilité que l'erreur, elle-même, soit détectée. Si le directeur ou la directrice prend sa décision en l'absence de données de recherche appropriées concernant le coût 3 (instruction de moins bonne qualité), il ou elle risque d'exagérer les coûts 1 et 2. Tout semble militer en faveur d'une approche conservatrice – c'est-à-dire conserver l'ancien livre pour une année supplémentaire. Cependant, nous pouvons essayer de mener une expérience avec les deux livres dans un modèle de théorie de la décision (Chernoff et Moses, 1959) et parvenir à une décision qui prenne explicitement en considération les divers coûts et probabilités. Le degré auquel les délibérations approfondies d'un excellent administrateur ou d'une excellente administratrice de l'éducation se rapprochent de ce modèle de théorie de la décision est un problème important qu'il convient d'étudier.

Facteurs compromettant la validité interne et externe

Dans les sections suivantes de ce chapitre, nous énonçons 12 facteurs compromettant la validité de divers protocoles de recherches expérimentales. Chaque facteur sera présenté dans le contexte des protocoles de recherche pour lesquels il représente un problème particulier. [...] Il importe de poser au préalable la distinction essentielle entre validité interne et validité externe. La validité interne est le minimum de base sans lequel une expérience est ininterprétable. Les traitements expérimentaux ont-ils réellement fait une différence dans ce

cas spécifique? La validité externe pose la question de la possibilité de *généraliser*. À quelles populations, à quels contextes, à quelles variables de traitement et de mesure cet effet peut-il être généralisé? Les deux types de critères sont évidemment importants, même s'ils sont souvent contradictoires dans la mesure où l'augmentation de l'un risque de compromettre l'autre. Tandis que la validité interne est la condition *sine qua non* et que la question de la validité externe, comme la question de l'inférence inductive, n'est jamais entièrement résolue, la sélection de protocoles de recherche donnant de bons résultats pour les deux types de validité est évidemment notre idéal. C'est particulièrement le cas de la recherche en éducation, où l'on vise de pouvoir généraliser à des cadres connus comme étant similaires. Tant les distinctions que les relations entre ces deux catégories de considérations sur la validité seront rendues plus explicites, au fur et à mesure qu'elles seront illustrées dans les discussions sur des protocoles de recherche spécifiques.

En ce qui concerne la validité interne, huit classes différentes de variables exogènes seront présentées. Ces variables, si elles ne sont pas contrôlées dans le plan expérimental, peuvent produire des effets confondus avec l'effet du stimulus expérimental. Ces variables correspondent à des effets :

1. Historique : les événements spécifiques survenus entre la première et la deuxième mesure, en plus de la variable expérimentale.

2. De maturation : renvoie aux processus qui, chez les répondants, s'accentuent avec le passage du temps, indépendamment des événements particuliers): par exemple vieillir, être de plus en plus affamé, de plus en plus fatigué, etc.

3. De test : les effets d'un test sur les résultats d'un second test.

4. D'instrumentation : des changements dans l'étalonnage d'un instrument de mesure ou des changements dans les observateurs ou les marqueurs utilisés peuvent entraîner des changements dans les mesures obtenues.

5. De régression statistique, lorsque les groupes ont été sélectionnés sur la base de leurs scores extrêmes.

6. De biais entraînant une sélection différentielle des répondants pour les groupes de comparaison.

7. De mortalité expérimentale, ou perte différentielle des répondant-e-s des groupes de comparaison.

8. De sélection, de maturation, d'interaction, etc. qui, dans certains des protocoles de recherches quasi expérimentales à groupes multiples, comme le protocole de recherche 10, sont confondues avec – c'est-à-dire, pourrait être pris à tort pour – l'effet de la variable expérimentale.

Concernant les facteurs affectant la validité externe, ou représentativité, nous discuterons des facteurs suivants :

9. L'effet réactif ou d'interaction du test : un pré-test peut augmenter ou diminuer la réactivité du répondant à la variable expérimentale. Cela peut rendre les résultats obtenus pour cette population pré-testée non représentatifs des effets pour l'univers non testé dans lequel les répondant-e-s expérimentaux ont été sélectionné-e-s.

10. Les effets d'interaction entre les biais de sélection et la variable expérimentale.

11. Les effets réactifs des arrangements expérimentaux qui empêcheraient la généralisation de l'effet de la variable expérimentale sur les personnes qui y sont exposées dans des environnements non expérimentaux.

12. L'interférence des traitements multiples, susceptible de se produire chaque fois que des traitements multiples sont appliqués aux mêmes répondant-e-s, parce que les effets des traitements antérieurs ne sont généralement pas effaçables.

Bibliographie

Brownlee, Kenneth A. 1960. *Statistical theory and methodology in Science and engineering*. New York: Wiley.

Campbell, Donald T. 1959. « Methodological suggestions from a comparative psychology of knowledge processes ». *Inquiry* 2 : 152-82.

Chernoff, Herman, et Lincoln E. Moses. 1959. *Elementary decision theory*. New York: Wiley.

Cox, David R. 1958. *Planning of experiments*. New York: Wiley.

Edward, Allen L. 1960. *Experimental design in psychological research*. New York: Rinehart.

Ferguson, George A. 1959. *Statistical analysis in psychology and education*. New York: McGraw-Hill.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Fisher, Ronald A. 1926. « The arrangement of field experiments ». *Journal of the Ministry of Agriculture* 33 : 503-15.

Fisher, Ronald A. 1935. *The Design of Experiments*. 1re éd. London: Oliver & Boyd.

Good, Carter V., et Scates. 1954. *Methods of research: Educational, Psychological, Sociological*. New York: Appleton-Century-Crofts.

Johnson, Palmer O. 1949. *Statistical methods in research*. New York: Prentice-Hall.

Johnson, Palmer O., et Robert W. B. Jackson. 1959. *Modern Statistical methods: Descriptive and inductive*. Chicago: Rand McNally.

Kendall, Maurice G., et William R. Buckland. 1957. A *dictionary of statistical terms*. London: Oliver & Boyd.

Lindquist, Everet F. 1953. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

McCall, William A. 1923. *How to Experiment in Education*. New York: Macmillan.

McNemar, Quinn. 1962. *Psychological statistics*. 3e éd. New York: Wiley.

Monroe, Walter S. 1938. « General Methods: Classroom experimentation » édité par G. M. Whipple. *Yearbook of the national society for the study of education* 2 : 319-27.

Roy, Samarenda N., et Ramanathan Gnanadesikan. 1959. « Some contributions to ANOVA in one or more dimensions: I and II ». *Annals of Mathematical Statistics* 30(2) : 304-317 ; 318-40. doi : 10.1214/aoms/1177706254.

Stanley, Julian C. 1957a. « Controlled experimentation in the classroom ». *Journal of Experimental Education*. 25:195-201.

Stanley, Julian C. 1957b. « Research methods: Experimental design ». *Review of Educational Research*. 27:449-59.

Stanley, Julian C. 1960. « Interactions of organisms with experimental variables as a key to the integration of organismic and variable-manipulating research », édité par E. M. Huddleston. *Yearbook of the National Council on Measurements used in Education*. 7-13.

Stanley, Julian C. 1961a. « Analysis of unreplicated three way classifications, with applications to rater bias and trait independence ». *Psychometrika* 26 : 205-20.

Stanley, Julian C. 1961b. « Studying status vs. manipulating variables ». in *Research design and analysis: The second Phi Delta Kappa symposium on educational research*, édité par R. O. Collier et S. M. Elam. Bloomington: Phi Delta Kappa, p. 173-208.

Stanley, Julian C. 1962. « Analysis-of-variance principles applied to the grading of essay tests ». *Journal of Experimental Education.* 30 : 279-83.

Stanley, Julian C., et D. E. Wiley. 1962. *Development and analysis of experimental designs for ratings*. Madison, Wis: Authors.

Thorndike, Edward L., William A. McCall et J. Crosby Chapman. 1916. « Ventilation in Relation to Mental Work ». *Teacher College, Columbia University Contributions to Education* (78).

Underwood, Benton J. 1957. *Psychological research.* New York: Appleton-Century-Crofts.

Winer, Ben J. 1962. *Statistical Principles in experimental design.* New York: McGraw Hill.

2. La méthode qualitative d'analyse d'impact

LAWRENCE B. MOHR

[Traduit de : Mohr, Lawrence B. 1999. « The qualitative method of impact analysis », *American Journal of Evaluation*, 20(1) : 69-84 (Extraits). Traduction par Carine Gazier et Valéry Ridde; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Le raisonnement causal

Il existe deux conceptualisations de la causalité, l'une factuelle et l'autre physique. Lorsque la recherche se déroule comme expliqué ci-dessus, l'inférence causale se fondera sur un argument qui justifie la raison opératoire et qui la lie au comportement, peut-être avec quelques réserves. Le type d'argument utilisé dans ce cas, de même que celui qui est utilisé dans des exemples qualitatifs antérieurs tels que l'établissement du lien entre la température des tuyaux et la surchauffe d'une voiture, sera appelé « raisonnement causal physique ». Lorsque l'inférence causale est une inférence de causalité factuelle, un autre type de processus de recherche est demandé, un processus de recherche aboutissant à la possibilité de faire valoir, encore peut-être avec des réserves, que la causalité factuelle a été établie. La forme de cet argument particulier serait appelée « raisonnement causal factuel ».

Le *raisonnement causal factuel* doit nécessairement s'appuyer sur une allégation contrefactuelle. Une bonne partie de l'énergie et de l'ingéniosité du protocole de recherche sera consacrée au fait de pouvoir argumenter de façon convaincante que si X ne s'était pas produit, Y non plus. Cela peut en réalité se faire dans le cadre d'une seule étude de cas,

ou lorsque l'on parle d'un seul événement qui n'est qu'une partie mineure d'une histoire plus vaste. En d'autres termes, on peut démontrer que X et Y se sont produits et soutenir autant que faire se peut, en se basant sur la connaissance des événements passés, en plus d'un examen minutieux et détaillé du contexte et des circonstances entourant l'événement en cours, que si X ne s'était pas produit ici, Y non plus. On ne pourrait pas faire délibérément une ou plusieurs observations empiriques d'un « groupe témoin » – impliquant un sujet et un contexte similaires dans lesquels X ne s'est pas produit. Néanmoins, il s'agirait d'un raisonnement causal factuel, d'un protocole de recherche quantitatif. Ou bien, comme le souligne le livre de King *et al.* (1994), on pourrait étudier en profondeur un ou deux cas dans lesquels X était présent et un ou deux cas où il ne l'était pas. Les auteurs ont tendance à considérer ces études à « petits N » comme qualitatives en raison de la profondeur de l'exploration et du mode de collecte et de mesure des données, mais dans la perspective du protocole de recherche, elles sont encore quantitatives. Toutefois, dans la plupart des cas, le raisonnement causal factuel est fondé sur des protocoles et des tailles d'échantillons pour lesquels des méthodes d'analyse statistique deviennent appropriées. Dans ce genre de cas, il est important de se rappeler, d'après la définition contrefactuelle, que le groupe témoin ou le groupe de comparaison ou la mesure « avant », par exemple, ne présente pas d'intérêt en soi. Il ne s'agit pas de comparer ce qui se passe (au temps présent) lorsque X se produit et ne se produit pas. Le groupe témoin ou de comparaison n'a d'intérêt que pour estimer ce qui se serait passé dans le groupe expérimental si X ne s'y était pas produit. C'est l'essence même du raisonnement causal factuel.

Prenons par exemple l'évaluation d'initiatives d'éducation à la santé. Il existe des protocoles de recherche statistiques de diverses sortes dans lesquels les chercheurs ont conclu que la *connaissance* de certains facteurs de risque entraîne (i.e. provoque) certains comportements liés à la santé, tels que des pratiques alimentaires plus saines sur le plan nutritionnel (Edwards, Acock, et Johnston, 1985). Souvent, on ne pense pas que cette connaissance des facteurs de risque puisse être une cause

physique. Le fait d'être exposé-e à des informations sur une alimentation correcte, par exemple, n'est pas considéré comme ce qui a *fait* manger correctement les individus, mais on constate plutôt que les gens suffisamment *motivés* (c'est ici qu'intervient la causalité physique) sont beaucoup plus susceptibles de manger correctement s'ils savent vraiment ce qui est correct que s'ils ne le savent pas – la connaissance précise est plus indiquée que l'ignorance ou la fausse idée. Quoi qu'il en soit, le raisonnement consiste à dire que celles et ceux qui ont reçu et absorbé le matériel d'éducation à la santé ont tendance à manger correctement. Cependant, celles et ceux qui ne l'ont pas reçu, qui ne l'ont pas compris ou qui ne s'en souvenaient pas ont eu tendance à ne pas manger correctement, ce qui signifie que la connaissance, au sein du premier groupe, était une condition nécessaire. (Mohr, 1995 : 267).

Le *raisonnement causal physique*, d'autre part, vise à démontrer de façon convaincante qu'un type particulier de lien physique s'est opéré entre deux événements dans le monde réel, essentiellement le lien entre une force et un mouvement. Il y a lieu de penser que c'est ce genre de raisonnement, par contradiction avec le raisonnement causal factuel, qui cible les causes *réelles*. En d'autres termes, fréquemment, nous ne sommes pas satisfait-e-s des preuves provenant d'études observationnelles ou même d'expériences aléatoires parce que le réel, c'est-à-dire le mécanisme causal physique, reste un mystère. Les preuves matérielles peuvent souvent être nécessaires à l'appui de la statistique, et elles contribueront presque toujours à un ensemble de données plus convaincantes si elles peuvent être obtenues.

Un exemple excellent est l'intérêt long et persistant porté à la question de savoir comment, physiologiquement, le tabagisme provoque le cancer du poumon. Nous avons été troublé-e-s, en d'autres termes, par le fait que la signature du tabac soit restée obscure. En revanche, il nous a fallu moins de temps pour nous convaincre de l'effet des fluorocarbures sur la couche d'ozone parce que nous disposions, non seulement des séries chronologiques statistiques reliant les deux, mais aussi d'un compte rendu des processus chimiques impliqués.

Pour revenir à l'évaluation de programme, considérons que nous pourrions essayer d'établir, s'il convient d'attribuer une baisse de la mortalité liée aux accidents de la route à une répression policière médiatisée des excès de vitesse dans une communauté donnée. Appliquer la méthode avant-après ou la méthode des séries chronologiques interrompues au taux de mortalité est une option. La comparaison avec une communauté qui est similaire et qui n'a pas fait l'objet d'une répression en est une autre. Ces deux méthodes sont classiques, quantitatives ou factuelles-causales, et toutes deux sont affaiblies par les menaces classiques pesant sur la validité des quasi-expériences. Une alternative à creuser à la place, ou peut-être mieux encore, en complément, nous amènerait à limiter notre attention à une seule communauté et à une seule période de préoccupation – une étude de cas en ce qui concerne cet aspect particulier du protocole. Nous pourrions utiliser une enquête pour déterminer, non seulement si et dans quelle mesure, mais aussi comment les citoyennes et citoyens ont pris conscience de l'existence de la répression. En d'autres termes, nous n'essaierions pas de montrer qu'elles et ils n'auraient jamais entendu parler de la répression sans ce projet médiatisant l'action policière (raisonnement causal factuel), mais plutôt de rendre compte du fait qu'elles et ils en ont effectivement pris connaissance par la lecture ou la télévision, ou qu'elles et ils en ont été informés par un policier, ou par un ami ou une amie qui a prétendu avoir vu le spot TV, et ainsi de suite – autant de processus physiques. De plus, nous chercherions à comprendre le comportement lui-même – pas seulement si les citoyen-ne-s conduisaient consciemment plus lentement, mais pourquoi – était-ce une perception accrue des dangers de l'excès de vitesse, par exemple, ou une probabilité subjective accrue de se faire prendre? Notez que la quantification de ces deux liens dans le projet, compte tenu de l'enquête en question, ne serait qu'une question de compter les citoyen-ne-s pour se faire une idée de la prévalence des processus causals découverts. La quantification ne ferait pas partie d'un protocole de recherche permettant d'en déduire la causalité. Le protocole de recherche est alors indubitablement qualitatif. Aucune preuve du contrefactuel ne serait

présentée, car l'enquête n'en contiendrait aucune. Au contraire, qu'un-e citoyen-ne ou qu'une centaine soit interrogé, le protocole de recherche dépend fondamentalement de la capacité à établir de manière convaincante l'existence des deux processus physiques de causalité chez les individus concernés – la connaissance de la répression en raison des activités du projet et la modification du comportement en raison des fortes motivations suscitées par le message du projet.

Un autre bref exemple pourrait être utile pour suggérer les types de rôles disponibles pour le raisonnement causal physique. Dans le type de recherche illustré ci-dessus par les pratiques alimentaires et la connaissance de la bonne nutrition, les efforts visant à accroître les connaissances seraient nécessairement évalués par le seul raisonnement causal factuel. Les connaissances nutritionnelles ne sont tout simplement pas une cause physique. Comme nous l'avons suggéré, il faudrait qu'interviennent des motivations ou des raisons distinctes, impulsant l'adoption du comportement alimentaire sain préconisé par le programme. Un programme plus approfondi chercherait également à inculquer ces motivations. Une analyse d'impact du programme pourrait alors être effectuée soit par un raisonnement factuel, soit par un raisonnement causal physique, soit par les deux. En d'autres termes, il n'y a aucune raison pour que les deux formes de logique ou de protocole de recherche ne puissent être prises en compte (a) dans le cadre d'une même évaluation, comme dans l'évaluation de l'initiative de diffusion de la connaissance par un raisonnement causal factuel, et des motivations par un raisonnement causal physique; ou (b) sur le même résultat, tant que nous nous soucions d'une cause physique, comme dans l'utilisation de deux protocoles de recherche différents pour déterminer si les individus ont changé ou non leurs habitudes alimentaires en raison des efforts déployés par le programme pour les convaincre des avantages qu'ils et elles pouvaient attendre de ce changement.

Cette contribution vise principalement à mettre en évidence l'idée d'un raisonnement causal physique et à suggérer ses fonctions. Les détails de la mise en œuvre réussie de ce raisonnement dans le cadre de l'évaluation

et d'autres recherches dépassent la portée de nos efforts ici, et ne se préciseront sans doute que grâce au cumul de plusieurs essais. Cependant, une idée importante concernant la mise en œuvre émerge de la logique définitionnelle de la causalité physique elle-même. Cette conception de la causalité repose sur la relation entre une force et un mouvement. Dans le comportement humain, ces éléments renvoient respectivement à la raison opérationnelle et au comportement lui-même, comme dans le rapport entre la peur de se faire prendre par la police et le respect de la limitation de vitesse.

Pour établir cette relation de façon convaincante, quatre éléments semblent importants. Il faut d'abord montrer que la personne ou le groupe étudié doit avoir eu la raison à laquelle nous voulons attribuer la causalité – qu'ils la revendiquent eux-mêmes ou non. Deuxièmement, il est important de démontrer que cette raison était très forte par rapport à d'autres considérations qui auraient pu être actives à l'époque. Troisièmement, il est généralement crucial de montrer par diverses autres manifestations que c'est effectivement cette raison qui opérait – en traçant la « signature » de la raison, dans le style de l'approche du *modus operandi*. Enfin, il sera généralement utile d'établir que l'on peut légitimement *interpréter* le comportement des personnes comme ayant pour intention de rester en dessous de la limite de vitesse, par exemple, ou de suivre une bonne pratique nutritionnelle, et pas seulement, au-delà du fait de simplement constater une conduite à une certaine vitesse ou la consommation certains aliments, comme le fait habituellement la recherche quantitative.

Ce serait une digression que de discuter longuement des divers autres points de vue sur la causalité et le raisonnement causal qui ont été suggérés par rapport à la recherche qualitative, comme le constructivisme, le néoréalisme, etc. Toutefois, il est peut-être bon de souligner un point, à savoir que ce qui est proposé ici, ce sont des *définitions* de la causalité, qui doivent normalement soutenir toutes ces perspectives. Que l'on adopte une vision constructiviste de la « causalité », ou que l'on mette l'accent sur les « réseaux de causalité »

plutôt que sur les dyades simples, linéaires, et de cause à effet, ou encore que l'on cherche avec le néoréaliste à découvrir des « mécanismes causaux » plutôt que des causes, on pourra toujours être confronté au défi de préciser ce qu'on entend réellement par « causalité » ou « causal » dans le contexte de l'approche choisie. Tout comme la méthode du *modus operandi*, ces autres perspectives ne dispensent pas de la nécessité de conceptualiser la causalité ni ne fournissent automatiquement elles-mêmes la définition conceptuelle requise. Elles ont tendance à utiliser le terme « cause » comme s'il avait été défini auparavant. Il est suggéré ici que la réponse à cette difficulté à définir puisse toujours être donnée en fonction de l'une des deux conceptualisations proposées ci-dessus. Il est donc suggéré, en outre, que les deux types de raisonnement causal, physique et factuel, seront en ce sens toujours pertinents, quel que soit le cadre philosophique, bien qu'ils ne soient peut-être pas toujours suffisamment importants pour l'argument de l'auteur pour mériter un intérêt. Enfin, bien que la recherche qualitative soit fermement établie comme une approche savante en sciences sociales et en évaluation de programmes, la prise en compte de la notion de causalité semble être régie par la définition contrefactuelle qui, du point de départ du protocole de recherche, est essentiellement quantitative. Ainsi, la possibilité de raisonner du point de vue de la causalité physique devrait ouvrir des perspectives de réflexion nettement plus larges, notamment dans le cadre des approches qualitatives.

Bibliographie

Edwards, Patricia K., Alan C. Acock, et Robert L. Johnston. 1985. « Nutrition behavior change: Outcomes of an educational approach ». *Evaluation Review* 9(4) : 441-60. doi : https://psycnet.apa.org/doi/10.1177/0193841X8500900404.

King, Gary, Robert O. Keohane et Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton: Princeton University Press.

Mohr, Lawrence B. 1995. *Impact analysis for program evaluation*. 2e éd. Newbury Park: Sage Publications.

3. Une évaluation sommative de la méthode expérimentale par assignation aléatoire, et une approche alternative de l'imputation causale

MICHAEL SCRIVEN

[Traduit de[1] : Scriven, Michael. 2008. « A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research ». *Journal of Multidisciplinary Evaluation*, 5(9) : 11-24 (Extraits). Traduction par Carine Gazier et Valéry Ridde. Article originellement publié en *open access*.]

A. *Le protocole expérimental par assignation aléatoire est une construction théorique d'un intérêt considérable. Cependant, il n'a essentiellement aucune application pratique lorsque cela concerne des êtres humains.* Il est important de préciser qu'une véritable étude expérimentale par assignation aléatoire doit être (au moins) en double aveugle, comme le sont toutes les bonnes études pharmacologiques, alors que les applications en matière de santé publique, d'éducation, de services sociaux, d'application de la loi, etc., qui sont actuellement préconisées dans les méthodes expérimentales par assignation aléatoire ne sont ni en double aveugle ni en simple aveugle, mais en « zéro-aveugle ». Les résultats de telles études sont donc susceptibles de s'expliquer en partie par un effet Hawthorne ou par son antithèse, puisqu'il est généralement

1. NdT : Cet article se concentre sur ce que l'on pourrait appeler un réexamen des références de la conception de la méthode expérimentale par assignation aléatoire, et propose quelques perspectives radicalement nouvelles à son égard. Sa lecture suppose une connaissance raisonnable des concepts de l'approche expérimentale. Les notes de bas de page ont été supprimées pour alléger le texte.

facile pour les membres des groupes expérimentaux et témoins de déterminer dans quel groupe ils et elles sont. *Par conséquent, l'argument commun selon lequel les modèles de la méthode expérimentale par assignation aléatoire préconisés dans des domaines tels que l'éducation, la santé publique, l'aide internationale, l'application de la loi, etc., ont l'avantage (unique) « d'éliminer toutes les explications fallacieuses » est totalement invalide.* Il était imprudent de supposer que l'assignation aléatoire des individus compenserait leur absence d'aveuglement (comme dans les études en simple aveugle), et encore moins l'absence de l'aveuglement des organisateurs et organisatrices des traitements, c'est-à-dire des prestataires de services (exigence qui distingue l'étude en double aveugle). Dans les sciences humaines appliquées, le label de méthode expérimentale par assignation aléatoire est en fait brandi par des pseudo-méthodes expérimentales par assignation aléatoire. Cet échec n'est pas dû à la négligence, mais à l'impossibilité presque totale, du moins dans les limites des protocoles habituels régissant l'expérimentation sur des sujets humains, d'organiser des conditions d'aveuglement, même en simple aveugle.

B. *Même les meilleures études en double aveugle sur les médicaments n'ont pas le pouvoir explicatif à caractère unique revendiqué par les partisanes et partisans de la méthode expérimentale par assignation aléatoire pour leurs études en zéro-aveugle.* Ces études, généralement considérées comme des exemples paradigmatiques du protocole expérimental par assignation aléatoire, sont elles-mêmes susceptibles d'être remises en question, car elles ne répondent pas aux exigences du protocole de recherche de la méthode expérimentale par assignation aléatoire. Cela a conduit à une demande pour ce que l'on appelle des études « triple aveugle ». Ce terme est défini de diverses manières et a été utilisé pour désigner l'aveuglement (ou l'exclusion par le biais d'expériences contrôlées par ordinateur) du/de la statisticien-ne qui analyse des résultats, du/de la pharmacien-ne lorsqu'il/elle administre des médicaments, ou encore du/de la radiologue ou pathologiste qui effectue la première étape de

l'interprétation des données. […] J'utiliserai ici le terme « totalement aveugle » pour traiter de tels cas. De plus, je le définirai comme une étude dans laquelle aucune personne impliquée dans la fourniture, l'obtention ou la réception du traitement, ou dans l'analyse des résultats, ne peut identifier l'appartenance d'un sujet à un groupe jusqu'à l'étape finale de décodage. […]

C. La difficulté de dissocier les effets « intrinsèques » d'un traitement, des effets psychologiques de l'administration du traitement, a deux résultats invalidants pour les inférences causales des protocoles de recherche expérimentaux. Le premier que nous venons d'évoquer – le problème selon lequel les effets distinctifs observés dans le groupe expérimental sont peut-être dus aux effets combinés de deux facteurs, et pas seulement aux effets du traitement. Le deuxième problème est qu'on ne peut pas dire quel sera l'effet du traitement expérimental s'il est administré en dehors du contexte expérimental, car on ne peut dire l'impact du contexte sur les effets. *Par conséquent, il est dangereux de généraliser l'utilisation, même dans le monde réel, du traitement expérimental mis à l'essai dans le cadre de la méthode expérimentale par assignation aléatoire. Toute étude expérimentale par assignation aléatoire digne de ce nom doit être complétée par de nombreux rapports de terrain – de grande qualité – sur les usages en situation réelle [hors cadre expérimental].*

D. Mais ce n'est pas tout. L'autre avantage logique revendiqué par les partisans de la méthode expérimentale par assignation aléatoire est qu'elle apporte la preuve de ce qu'on dit être la principale propriété logique des causes : le fait qu'elles résistent à l'épreuve du contrefactuel (c'est-à-dire qu'une cause est quelque chose sans laquelle l'effet ne se serait pas produit). *Cependant, cette propriété de soutien contrefactuel n'est certainement pas une propriété logique des causes, […] et encore moins*

une propriété logique que possèdent les quasi-méthodes expérimentales par assignation aléatoire [...]. Ce n'est pas une propriété logique des causes, en l'état, à cause du phénomène commun de la surdétermination, c'est-à-dire des situations où un effet E est causé par un événement C, mais où E se serait produit même si C ne s'était pas produit, à cause de la situation D, qui est « cachée dans le décor », prête à passer à l'acte si C ne le fait pas. Par exemple, lorsqu'un joueur de baseball – appelons-le Jaime Cortez – attrape facilement une balle dans un match professionnel (affirmation causale), le fait qu'un deuxième joueur se soit positionné derrière lui et aurait attrapé la balle si Cortez l'avait manquée, ne nous amène pas à dire que Cortez n'a pas attrapé la balle, bien qu'il soit clair que le contrefactuel ne tienne pas. [...]

E. *Les menaces que font peser les variables confondantes sur les protocoles expérimentaux par assignation aléatoire sont extrêmement graves et nombreuses, et leur traitement est coûteux. Elles exigent une attention continue et la mobilisation de compétences importantes, qui excèdent souvent les ressources disponibles dans le cadre de ces études.* [...]

F. Il ressort de l'examen des sections précédentes que : (i) le potentiel régulièrement revendiqué par les méthodes expérimentales par assignation aléatoire ne se retrouve pas dans les quasi-méthodes expérimentales par assignation aléatoire du monde réel; et que : (ii) ces protocoles de recherche du monde réel ont des handicaps sérieux qui leur sont propres. Est-ce que la démarche d'assignation aléatoire leur laisse malgré tout une supériorité générale résiduelle? En d'autres termes, est-ce que les quatre limites énumérées plus haut – les écarts entre elles et la méthode expérimentale par assignation aléatoire idéale – leur empêche d'exclure toute cause autre que le traitement expérimental? Fondamentalement, la réponse est à la fois oui – dans un sens (de

« spécial ») – et non, dans un autre sens; et les deux sens s'annulent l'un l'autre, laissant la quasi-méthode expérimentale par assignation aléatoire sans avantage net. […]

G. *La véritable « référence » pour les affirmations causales est la même norme ultime que pour toutes les affirmations scientifiques; il s'agit de l'observation critique.* La causalité peut être observée directement, en laboratoire, à la maison ou sur le terrain, comme l'une des nombreuses observations que l'on peut faire, toujours indexées à un contexte, à l'instar de la fonte du plomb obtenue en chauffant un creuset, la friture des œufs dans une poêle ou le faucon emmenant un pigeon. La causalité peut aussi être déduite d'observations directes non causales, sans expérimentation, comme par exemple un médecin légiste qui effectue une autopsie pour déterminer la cause du décès. […]

H. La première implication [de ce qui précède] est que *l'étude de cas réalisée par des professionnel-le-s, dans la mesure où elle est souvent imprégnée d'affirmations causales fondées sur l'observation, peut à nouveau prétendre à une capacité honorable de démonstration de la causalité.* Soyons clairs sur un point, nous ne faisons référence qu'aux meilleurs exemples de ce genre. Par le passé, un grand nombre de travaux désespérément peu scientifiques ont été présentés sous le nom de « méthodes qualitatives », y compris de nombreux rapports anecdotiques décrits comme des études de cas, et le présent argument ne les réhabilite pas. Nous voulons simplement éviter la culpabilité par association, et permettre qu'il y ait des études de cas réalisées dans la tradition critique du bon travail scientifique où ce qui est rapporté est vérifié et inclut des affirmations causales. […]

I. *Le point essentiel à ce stade de la discussion est le suivant : suggérer qu'on ne pourrait établir de causalité en évaluation sans la démarche expérimentale par assignation aléatoire est aussi farfelu que de suggérer qu'on ne pourrait établir de causalité en Histoire sans ce type de méthode (on ne pourrait alors, faute d'expérimentation, établir par exemple que la guerre en Irak a causé la mort de nombreux citoyens États-Uniens) […].* Presque toutes les affirmations causales formulées dans le monde réel, qui sont hors de tout doute raisonnable, sont fondées sur l'observation ou l'inférence directe de l'observation, que ce soit dans le contexte d'un laboratoire scientifique, d'une clinique ou d'un travail de terrain, ou dans la pratique du droit ou de l'histoire, ou dans les affaires courantes ou le journalisme. Elles peuvent être rassemblées et analysées statistiquement, qu'il s'agisse de la focale microscopique des études de cas ou du regard plus surplombant des études à grande échelle […].

J. Bien sûr, il existe des programmes, dans le cadre de l'aide internationale et dans d'autres domaines, pour lesquels il n'est pas si facile d'établir qu'ils produisent des avantages ni de déterminer l'ampleur de ceux-ci. Comme indiqué précédemment, il s'agit probablement de programmes dont les effets sont moins importants, moins immédiats et moins évidents, en particulier des programmes dont les effets nets ne sont tout simplement pas observables dans chaque cas. Pour certains d'entre eux, mais pas tous, un protocole expérimental par assignation aléatoire peut être approprié. Pour d'autres – probablement la majorité, mais personne n'a fait le compte – d'autres protocoles de recherche sont plus pertinents. Il est maintenant temps de se demander s'il existe une méthode sous-jacente qui peut être utilisée ou qui sous-tend logiquement toutes les affirmations causales légitimes, puisqu'il ne s'agit manifestement pas de l'expérimentation par assignation aléatoire. Cela nous servira non seulement pour choisir des protocoles de recherche plus spécifiques, mais dans de nombreux cas où il n'est pas nécessaire de concevoir des protocoles plus spécifiques, et dans ceux où aucun d'entre eux ne fonctionne. […]

K. Le coup de grâce contre la méthode expérimentale par assignation aléatoire, dans les guerres causales, vient d'un contrôle d'authenticité. Le critère ultime de l'authenticité est l'auto-application, le cas échéant. Par exemple, il est souvent pertinent que les évaluateurs ou évaluatrices fassent évaluer leur propre travail, et un test d'authenticité consiste à voir à quelle fréquence ils ou elles le font et à s'assurer que cela est fait avec le soin qu'ils ou elles demandent dans leurs propres appels aux autres pour que leurs programmes soient évalués. À présent, si la cause de la méthode expérimentale par assignation aléatoire est légitime, ne serait-ce pas une bonne pratique, pour celles et ceux qui en font un usage exclusif, que de vérifier si leur politique définie sur cette base fonctionne? En d'autres termes, d'évaluer leur propre politique. S'ils et elles ne le font pas, ne peut-on pas conclure qu'ils et elles échouent à un test crucial de leur propre doctrine?

4. L'utilisation des méthodes qualitatives pour l'explication causale

JOSEPH A. MAXWELL

[Traduit de : Maxwell, Joseph A. 2004. « Using Qualitative Methods for Causal Explanation ». *Fields Methods*, 16 : 246-251 (Extraits). Traduction par Carine Gazier et Valéry Ridde; traduction et reproduction du texte avec l'autorisation de Sage Publications.]

Une approche réaliste de l'explication causale

La compréhension philosophique de la causalité a connu un changement significatif au cours des cinquante dernières années (Salmon, 1998), changement qui n'a pas été pleinement apprécié par de nombreux spécialistes des sciences sociales. Ce changement est en grande partie le résultat de l'émergence du réalisme comme alternative au positivisme/empirisme et au constructivisme en tant que philosophie de la science (Archer *et al.*, 1998; Baert, 1998; Layder, 1990; Pawson et Tilley, 1997; Putnam, 1990; Sayer, 1992).

Les « réalistes » appréhendent généralement la causalité comme étant constituée non pas de régularités, mais de mécanismes et de processus causaux réels (et en principe observables), qui peuvent produire ou non des régularités. Pour la philosophie de la science en général, cette approche de la causalité a été élaborée de façon plus systématique par Salmon (1984, 1998). Pour les sciences sociales, elle est souvent associée (mais pas seulement) à celles et ceux qui se disent « réalistes critiques » (Archer *et al.*, 1998; Sayer, 1992). La critique adressée par le réalisme critique à la conception de la causalité fondée sur les régularités a remis en question, non seulement le fait de limiter notre connaissance de la

causalité aux régularités observées, mais aussi le fait de négliger les influences contextuelles (Pawson et Tilley, 1997; Sayer, 1992) et les processus mentaux (Davidson, 1980, 1993; McGinn, 1991) comme faisant partie intégrante de l'explication causale dans les sciences sociales, et de nier que nous puissions observer directement la causalité dans des instances spécifiques (Davidson, 1980; Salmon, 1998).

Cette vision réaliste de la causalité est en cohérence avec les caractéristiques essentielles de la recherche qualitative, y compris celles mises en avant par les constructivistes. Premièrement, l'affirmation selon laquelle certains processus causaux peuvent être directement observés, plutôt que seulement déduits de la co-variation mesurée des causes et des effets présumés, renforce l'importance accordée par de nombreux chercheurs qualitatifs à l'observation et à l'interprétation directes des processus sociaux et psychologiques. Si une telle observation directe est possible, alors elle peut l'être dans des cas isolés plutôt que d'exiger une comparaison des situations dans lesquelles la cause présumée est présente ou absente. Cela confirme la valeur des études de cas pour l'explication causale. Deuxièmement, en considérant que le contexte est intrinsèquement impliqué dans les processus causaux, elle rejoint l'insistance des chercheurs et chercheuses qualitativistes sur l'importance explicative du contexte et le fait d'une manière qui ne se contente pas de réduire ce contexte à un ensemble de « variables exogènes ». Troisièmement, l'argument réaliste selon lequel les événements et les processus mentaux sont des phénomènes réels qui peuvent être des causes de comportement, converge avec le rôle fondamental que les chercheuses et chercheurs qualitativistes attribuent au sens et à l'intention dans l'explication des phénomènes sociaux et la nature essentiellement interprétative de notre compréhension de ces derniers (Maxwell, 1999, 2004a). Quatrièmement, en affirmant que l'explication causale ne dépend pas intrinsèquement de comparaisons préétablies, elle légitime l'utilisation par les chercheuses et chercheurs qualitativistes de conceptions et de méthodes souples et inductives.

Le réalisme est également compatible avec de nombreuses autres caractéristiques du constructivisme et du postmodernisme (Baert, 1998 : 174; Maxwell, 1995, 1999, 2004b), y compris l'idée que la différence est fondamentale plutôt que superficielle, un scepticisme à l'égard des « lois générales », de l'anti-fondationnalisme et d'une épistémologie relativiste. Lorsqu'elle diffère de celles-ci, c'est principalement dans son ontologie réaliste – impliquant qu'il existe un monde réel, même s'il n'est pas « objectivement » connaissable – et par son accent sur la causalité (bien qu'un concept fondamentalement différent de la causalité sociale que celui des positivistes). Putnam (1990), l'une des figures majeures du développement du réalisme, affirme :

> Que la causalité « existe vraiment » ou non, elle existe certainement dans notre « monde réel ». Ce qui la rend réelle au sens phénoménologique du terme, c'est la possibilité de demander « Est-ce vraiment la cause? », c'est-à-dire de vérifier les affirmations causales, d'apporter de nouvelles données et de nouvelles théories à leur sujet… Le monde du langage ordinaire (le monde dans lequel nous vivons réellement) est plein de causes et d'effets. Ce n'est que lorsque nous insistons sur le fait que le monde du langage ordinaire (ou le *Lebenswelt*) est défectueux… et que nous cherchons un monde « véritable » … que nous finissons par nous sentir obligé-e-s de choisir entre l'image « d'un univers physique avec une structure intégrée » et celle « d'un univers physique avec une structure imposée par l'esprit ». (Putnam, 1990 : 89)

Théorie de la variance et théorie des processus comme formes d'explication causale

La distinction philosophique entre les approches positivistes/empiristes et réalistes de la causalité est remarquablement semblable à la distinction qui existe par ailleurs entre deux approches de recherche, que Mohr (1982, 1995, 1996) qualifie de théorie de la variance et de théorie des processus. La théorie de la variance traite des variables et des corrélations entre elles. Elle est basée sur une analyse de la contribution des différences de valeurs de certaines variables aux différences d'autres variables. La théorie de la variance, qui implique idéalement une mesure précise des différences et des corrélations, tend à être associée à la recherche qui utilise l'échantillonnage probabiliste, la mesure quantitative, la vérification statistique des hypothèses et les protocoles de recherche expérimentaux ou corrélatifs. Comme le note Mohr, « l'interprétation archétype de cette idée de causalité est le modèle de régression linéaire ou non linéaire » (Mohr, 1982 : 42).

La théorie des processus, en revanche, traite des événements et des processus qui les concernent. Elle est fondée sur une analyse des processus de causalité par lesquels certains événements en influencent d'autres. L'explication des processus, puisqu'elle traite d'événements et de processus spécifiques, se prête moins bien aux approches statistiques. Elle se prête à l'étude approfondie d'un ou de quelques cas, ou d'un échantillon relativement restreint d'individus et à des formes textuelles de données qui conservent les liens chronologiques et contextuels entre les événements.

Des distinctions similaires entre les approches de variance et de processus dans les sciences sociales sont celles qui existent entre « l'analyse des variables » et le « processus d'interprétation » (Blumer, 1956), les approches axées sur les variables et les cas (Ragin, 1987) et les théories factorielles et explicatives (Yin, 1993 : 15). Gould (1989) décrit deux approches en sciences naturelles. La première est caractéristique

de la physique et de la chimie, domaines qui reposent sur des méthodes expérimentales et font appel aux lois générales, tandis que la seconde est caractéristique de disciplines telles que la biologie évolutive, la géologie et la paléontologie, qui traitent de situations uniques et de séquences historiques. Il affirme que

> La résolution de l'histoire doit être enracinée dans la reconstruction des événements passés eux-mêmes – selon leurs propres termes – sur la base de preuves narratives de leurs propres phénomènes uniques… La science historique n'est pas pire, plus restreinte ou moins capable d'aboutir à des conclusions fermes dès lors qu'elle ne procède pas par l'expérimentation, la prédiction et la substitution (subsomption) sous les lois invariantes de la nature. Les sciences de l'histoire utilisent un mode d'explication différent, enraciné dans la richesse comparative et observationnelle de nos données. (Gould, 1989 : 277-279)

Les deux types de théories impliquent une explication causale. La théorie des processus n'est pas simplement « descriptive », par opposition à la théorie des variances « explicatives ». Il s'agit d'une approche différente de l'explication. Les méthodes expérimentales et les méthodes quantitatives impliquent généralement une approche du problème de causalité par la « boîte noire ». Faute d'informations directes sur les processus sociaux et cognitifs, elles doivent tenter de corréler les différences de résultats avec des différences entre les intrants et de contrôler d'autres facteurs susceptibles d'influer sur les résultats. Les méthodes qualitatives, en revanche, peuvent souvent examiner directement ces processus causaux, bien que la validité de leurs conclusions soit sujette à des menaces qui leur sont propres.

Un exemple frappant de la différence entre les approches de variance et de processus est un débat qui a eu lieu dans la *New York Review of Books*, sur la validité scientifique de la psychanalyse. Crews (1993) et Grünbaum

(1994) ont nié que la psychanalyse soit scientifique parce qu'elle ne répond pas aux critères scientifiques de vérification, critères que même les explications psychologiques plus courantes doivent satisfaire :

> Pour justifier qu'un facteur X (comme le fait d'être insulté) ait un lien causal pertinent avec une sorte de résultat Y (comme le fait d'être en colère ou de se sentir humilié) dans une classe de référence C, il est nécessaire de prouver que l'incidence de Y dans la sous-classe des X est *différente* de son incidence dans la sous-classe des non-X… En l'absence de telles statistiques, il n'y a manifestement pas de raison suffisante d'attribuer l'oubli des expériences négatives à leur mécontentement affectif, et encore moins d'attribuer les symptômes névrotiques à la répression de ces expériences. (Grünbaum, 1994 : 54; mis en exergue dans l'original)

Nagel (1994a, 1994b) a convenu avec Grünbaum que les explications générales de Freud à l'égard de nombreux phénomènes psychologiques sont suspectes, mais considère la principale contribution de Freud, non pas comme la promulgation d'une telle théorie générale, mais comme le développement d'une méthode de compréhension basée sur des interprétations et des explications individuelles. Il a également convenu « que les hypothèses psychanalytiques sont causales et nécessitent une confirmation empirique; mais nous ne sommes pas d'accord sur le type de preuve qui compte le plus » (Nagel, 1994b : 56). Le type d'explication que Nagel a défendu comme caractéristique, à la fois de la psychologie du sens commun et de la psychanalyse, implique une compréhension spécifique de cas particuliers fondée sur un cadre d'interprétation général, une compréhension fondée sur le « rapprochement » d'éléments de preuve d'une manière qui puisse expliquer comment un résultat particulier s'est produit plutôt que de démontrer l'existence d'une relation statistique entre des variables spécifiques.

Les chercheurs et chercheuses qualitativistes ont fourni de nombreuses illustrations de la façon dont une telle approche par les processus peut être utilisée pour élaborer des explications causales. Par exemple, Weiss (1994) soutient que :

> Dans les études par entretiens qualitatifs, la démonstration du lien de causalité repose fortement sur la description d'une séquence d'événements visualisables, chaque événement se succédant à l'autre… Des études quantitatives soutiennent l'affirmation du lien de causalité en montrant une corrélation entre un événement antérieur et un événement ultérieur. Une analyse des données recueillies dans le cadre d'une enquête par sondage à grande échelle pourrait, par exemple, montrer qu'il existe une corrélation entre le niveau de diplôme de l'épouse et le caractère égalitaire du mariage. Dans les études qualitatives, nous nous intéressons aux processus par lesquels le niveau de diplôme de l'épouse ou les facteurs associés à son éducation s'expriment dans l'interaction conjugale. (Weiss, 1944 : 179)

Un deuxième exemple est fourni par une étude en méthodes mixtes sur les chutes de patient-e-s dans un hôpital (Morse et Tylko, 1985; Morse, Tylko, et Dixon, 1987). Elle comprenait des observations qualitatives et des entretiens avec des patient-e-s âgé-e-s qui étaient tombé-e-s, revenant sur la façon dont elles et ils se déplaçaient dans l'environnement hospitalier et sur les raisons de leur chute. Les chercheurs et chercheuses ont utilisé ces données pour déterminer les causes des chutes, telles que l'utilisation de meubles ou de poteaux pour intraveineuses comme supports, qui n'avaient pas été rapportées dans des études quantitatives antérieures. Cette identification a été rendue possible par le fait que l'étude a mis en exergue le processus de déambulation des patient-e-s ainsi que les événements et les circonstances spécifiques qui ont mené à la chute, plutôt que la recherche d'une corrélation entre les chutes et d'autres variables précédemment définies.

Toutefois, l'élaboration d'explications causales dans une étude qualitative n'est pas une tâche simple. En outre, la validité de toute explication causale est menacée par de nombreux facteurs potentiels, qui devront être pris en compte lors de la conception et la réalisation de toute étude. À cet égard, la situation de la recherche qualitative ne diffère pas de celle de la recherche quantitative. Les deux approches doivent identifier et traiter les menaces susceptibles de peser sur la validité de toute explication causale proposée. Cette capacité à exclure des explications différentes plausibles ou « hypothèses rivales », plutôt que l'utilisation de méthodes ou de modèles spécifiques, est largement considérée comme la caractéristique fondamentale de la recherche scientifique en général (Campbell, 1986 : 125; Platt, 1966; Popper, 1959).

Bibliographie

Archer, Margaret, Roy Bhaskar, Andrew Collier et Alan Norrie. 1998. *Critical Realism: Essential readings*. London: Routledge.

Baert, Patrick. 1998. *Social Theory in the Twentieth century*. New York: New York University Press.

Blumer, Herbert. 1956. « Sociological Analysis and the 'variable' ». *American Sociological Review* 21(6) : 683-90.

Campbell, Donald T. 1986. « Science's social system of validity-enhancing collective belief change and the problems of the social sciences ». in *Metatheory in social science: Pluralisms and subjectivities*, édité par D. W. Fiske et R. A. Shweder. Chicago: University of Chicago Press, p. 108-35.

Crews, Frederick C. 1993. « The unknown Freud ». *New York Review of Books*.

Davidson, Donald. 1980. *Essays on actions and events*. Oxford: Clarendon.

Davidson, Donald. 1993. « Thinking Causes ». in *Mental causation*, édité par J. Heil et A. Mele. Oxford: Clarendon, p. 3-17.

Gould, Stephen J. 1989. *Wonderful life: The Burgess Shale and the nature of history*. New York: W. W. Norton.

Grünbaum, Adolf. 1994. « Freud's permanent revolution: An Exchange ». *New York Review of Books* 54-55.

Layder, Derek. 1990. *The realist image in social science*. New York: St. Martins.

Maxwell, Joseph A. 1995. « Diversity and methodology in a changing world ». *Pedagogia* 30 : 32-40.

Maxwell, Joseph A. 1999. « A realist/postmodern concept of culture ». in *Anthropological theory in North America*, édité par E. L. Cerroni-Long. Westport : Bergin & Garvey, p. 143-73.

Maxwell, Joseph A. 2004a. « Causal explanation, qualitative research, and scientific inquiry in education ». *Educational Researcher* 33(2) : 3-11.

Maxwell, Joseph A. 2004b. « Re-emergent scientism, postmodernism, and dialogue across differences ». *Special issue of Qualitative Inquiry* 10(1) : 35-41.

McGinn, Colin. 1991. « Conceptual causation: Some elementary reflections ». *Mind* 100 4 : 573-86.

Mohr, Lawrence B. 1982. *Explaining organizational behavior*. San Francisco: Jossey-Bass.

Mohr, Lawrence B. 1995. *Impact analysis for program evaluation*. 2e éd. Thousand Oaks: Sage Publications.

Mohr, Lawrence B. 1996. *The causes of human behavior: Implications for theory and method in the social sciences*. Ann Arbor: University of Michigan Press.

Morse, Janice M., et Suzanne J. Tylko. 1985. « The use of qualitative methods in a study examining patient falls ». Washington, DC.

Morse, Janice M., Suzanne J. Tylko, et Herbert A. Dixon. 1987. « Characteristics of the fall-prone patient ». *The Gerontologist* 27(4) : 516-22.

Nagel, Thomas. 1994a. « Freud's Permanent Revolution ». *New York Review of Books* 34-38.

Nagel, Thomas. 1994b. « Freud's Permanent Revolution: An Exchange ». *New York Review of Books* 55-56.

Pawson, Ray, et Nicholas Tilley. 1997. *Realistic evaluation*. London: Sage Publications.

Platt, John R. 1966. « Strong inference ». *Science* 146 : 347-53.

Popper, Karl. 1959. *The logic of scientific discovery*. New York: Basic Books.

Putnam, Hilary. 1990. *Realism with a human face*. Cambridge: Harvard University Press.

Ragin, Charles C. 1987. *The comparative method : Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Salmon, Wesley C. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, Wesley C. 1998. *Causality and explanation*. New York : Oxford University Press.

Sayer, Andrew. 1992. *Method in social science: A realist approach*. 2e éd. London: Routledge.

Weiss, Robert S. 1994. *Learning from strangers: The art and method of qualitative interviewing*. New York: Free Press.

Yin, Robert K. 1993. *Applications of case study research*. Thousand Oaks: Sage Publications.

5. Trois étapes pour construire et tester des théories de moyenne portée dans le cadre d'essais contrôlés randomisés réalistes : les leçons théoriques et méthodologiques d'une application

FARAH JAMAL, ADAM FLETCHER, NICHOLA SHACKLETON, DIANA ELBOURNE, RUSSELL VINER ET CHRIS BONELL

[Traduit de : Jamal, Farah, Adam Fletcher, Nichola Shackleton, Diana Elbourne, Russell Viner et Chris Bonell. 2015. « The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example ». *Trials*, 16 (466) : 1-10 (Extraits). Traduction par Carine Gazier et Valéry Ridde. Article originellement publié en *open access*.]

La plupart des grands défis de santé publique actuels sont complexes et nécessitent des interventions pour faire face à de multiples déterminants aux niveaux individuel et socioécologique (Choksi et Farley, 2012; Craig *et al.*, 2008). Les essais contrôlés randomisés (ECR, ou méthode expérimentale par assignation aléatoire), bien conçus et réalisés de manière appropriée, ont la meilleure validité interne pour estimer les effets d'interventions complexes afin de déterminer si des interventions spécifiques sont efficaces ou non dans leur ensemble. Cependant, on reproche souvent aux méthodes expérimentales par assignation aléatoire de ne pas ouvrir la « boîte noire » – c'est-à-dire d'examiner à gros traits

« ce qui fonctionne » sans expliquer les processus sous-jacents et les mécanismes d'action, et comment ils varient selon le contexte et les caractéristiques de la personne et du lieu.

Bien que certain-e-s auteurs et autrices aient affirmé que des effets de traitement de faibles ampleurs, mais importants pour l'évaluation seraient assez extrapolables dans tous les contextes (Peto, Collins, et Gray, 1995), cela ne concerne que les essais pharmaceutiques. Les interventions de santé publique, en revanche, sont loin d'être directement extrapolables, car elles impliquent l'interaction complexe de la structure et de l'agencéité (*agency*), façonnant la mise en œuvre des interventions et la manière dont les gens y réagissent, laquelle variera considérablement selon les contextes nationaux et locaux. Ainsi, elles peuvent être conceptualisées comme des « interruptions » de systèmes sociaux complexes (Hawe, Shiell, et Riley, 2009). Pour que les résultats des évaluations en santé publique soient utiles pour décider si les interventions doivent être déployées ailleurs, il faut davantage tenir compte de la validité externe des résultats de l'évaluation. Pour cela, il faut, non seulement déterminer si les interventions peuvent être mises en œuvre dans un nouveau cadre, mais aussi comprendre clairement les mécanismes de mise en œuvre et de causalité des interventions et la manière dont chacun de ces mécanismes peut varier en fonction du contexte (Cartwright et Hardie, 2012).

Une approche réaliste de l'expérimentation par assignation aléatoire

Des essais contrôlés randomisés réalistes ont été proposés comme approche d'évaluation scientifique permettant de combler ces lacunes tout en préservant les points forts des méthodes expérimentales par assignation aléatoire, qui fournissent des données probantes ayant une grande validité interne dans l'estimation des effets (Bonell *et al.*, 2012;

Jamal *et al.*, 2013). Les évaluateurs et évaluatrices réalistes ont considéré que les interventions « fonctionnent » en introduisant des *mécanismes* qui interagissent avec les caractéristiques de leur *contexte* pour produire des *résultats*, et ils posent et répondent à des questions, non seulement sur ce qui fonctionne à un niveau global, mais aussi sur *ce qui fonctionne pour qui et dans quelles circonstances* (Pawson et Tilley, 1997).

Un aspect essentiel de l'évaluation réaliste consiste à comprendre le fonctionnement de l'intervention en anticipant la diversité des mécanismes d'intervention potentiels, en les présentant dans le cadre d'une théorie du changement, et en évaluant empiriquement, si et comment ces mécanismes sont « activés » ou « entravés » dans les différents contextes dans lesquels l'intervention est réalisée, et comment cela peut varier selon les groupes. Le contexte fait référence à l'ensemble préexistant de situations sociales, de normes, de valeurs et d'interdépendances (par exemple, structure organisationnelle, situation géographique, caractéristiques démographiques des participantes et participants) au sein duquel une intervention est mise en œuvre. L'évaluateur ou l'évaluatrice doit donc émettre des hypothèses et vérifier comment la théorie du changement portée par l'intervention interagit avec le contexte pour permettre (ou empêcher) la mise en œuvre, les mécanismes causaux et, en fin de compte, les résultats. Pawson et Tilley (1997) décrivent ce processus d'élaboration d'hypothèses comme le développement de « configurations contexte-mécanisme-résultat (CMR) ». Traditionnellement, les évaluateurs et évaluatrices réalistes examinent ces hypothèses à l'aide de données d'observation, rejetant l'utilisation de groupes de répartition aléatoire et de groupes témoins comme étant l'incarnation d'une épistémologie positiviste aux antipodes de leur propre orientation réaliste.

Les partisanes et partisans des essais randomisés réalistes partagent l'idée que les évaluations traditionnelles ont été trop axées sur des questions d'effets globaux, ainsi que la nécessité de théoriser et d'évaluer empiriquement la façon dont les mécanismes d'intervention interagissent avec le contexte pour produire des résultats. Cependant, elles et ils

reconnaissent également qu'un groupe de comparaison réparti de façon aléatoire est le moyen le moins biaisé pour déterminer l'orientation et l'ampleur des effets d'une intervention et la façon dont ils sont modérés par le contexte, en soulignant que les études observationnelles sont souvent entravées par l'absence de point de comparaison au départ. Les méthodes expérimentales par assignation aléatoire réalistes sont donc un protocole de recherche qui permet de centrer les évaluations sur le perfectionnement de la théorie d'intervention extrapolable, tout en évaluant si des interventions particulières ont été efficaces ou non, puisque les deux questions peuvent être examinées dans le cadre de modèles de méthodes expérimentales par assignation aléatoire modifiés (Bonell *et al.*, 2012; Jamal *et al.*, 2013; Moore *et al.*, 2015). Les partisan-e-s de l'approche réaliste de la méthode expérimentale par assignation aléatoire rejettent l'idée que la méthode expérimentale par assignation aléatoire serait nécessairement positiviste, en invoquant les écrits scientifiques selon lesquels la recherche ne peut, en pratique, être réduite à des paradigmes discrets et incommensurables et que les méthodes n'impliquent pas nécessairement des épistémologies (Bonell *et al.*, 2013). Les méthodes expérimentales par assignation aléatoire réalistes se distinguent également des évaluations réalistes imbriquées au sein des méthodes expérimentales par assignation aléatoire, mais qui se déroulent en parallèle sans utiliser les comparaisons assignées aléatoirement comme ressource analytique (Byng *et al.*, 2008).

Alors que certains initiateurs et certaines initiatrices de l'évaluation réaliste rejettent l'argument selon lequel l'analyse réaliste peut s'appuyer sur des données expérimentales (Marchal *et al.*, 2013), la nouvelle décision du Conseil de la recherche médicale (CRM) sur l'évaluation des processus (Moore *et al.*, 2014, 2015) a reconnu la valeur des méthodes expérimentales par assignation aléatoire réalistes pour poser une série de questions sur la mise en œuvre, le contexte, les mécanismes, les résultats et la normalisation. Toutefois, l'approche réaliste de la méthode expérimentale par assignation aléatoire n'a jusqu'à présent été décrite que dans la théorie (Bonell *et al.*, 2012, 2013) et aucun exemple détaillé n'a été fourni.

Dans le contexte de l'intérêt croissant porté à la conception et à la conduite de méthodes expérimentales par assignation aléatoire réalistes, il est urgent de fournir des conseils pratiques sur la manière dont une telle approche peut être mise en œuvre.

Cet article fournit un exemple concret du processus théorique et méthodologique de réalisation d'une méthode expérimentale par assignation aléatoire réaliste s'appuyant sur l'évaluation de l'intervention « apprendre ensemble » (AE). Il s'agit d'un regroupement de méthodes expérimentales par assignation aléatoire de trois ans en « phase III » (Craig *et al.*, 2008) portant sur 40 établissements (20 dans le groupe d'intervention et 20 dans le groupe témoin) en Angleterre, afin d'évaluer une approche visant, à l'échelle de l'établissement, à réduire les situations de harcèlement et de violences (Bonell *et al.*, 2014). En s'appuyant sur un essai pilote réalisé dans huit établissements au cours d'une année scolaire (2011-2012) (Bonell *et al.*, 2015), la phase III de la méthode expérimentale par assignation aléatoire a débuté en septembre 2014 et est en cours (en 2015). Le protocole a déjà été publié (Bonell *et al.*, 2014).

À partir de cet exemple, cet article décrit les étapes par lesquelles une méthode expérimentale par assignation aléatoire réaliste développe et teste des hypothèses sur la façon dont les mécanismes d'intervention interagissent avec divers contextes pour produire des résultats. Le protocole d'essai publié n'aborde pas ces questions et est axé sur les méthodes d'évaluation des effets globaux et de la fidélité des interventions. Le présent article ne s'écarte pas de ce protocole (Bonell *et al.*, 2014), mais l'étend afin de fournir un exemple concret aux chercheurs et chercheuses, aux praticien-ne-s et aux décideurs et décideuses politiques qui cherchent à concevoir des méthodes expérimentales par assignation aléatoire réalistes, bien qu'il ne fournisse pas de résultats d'études puisqu'ils ne sont pas encore disponibles. Nous discutons également des conséquences importantes pour celles et ceux qui participent à l'établissement de rapports et à l'examen de méthodes expérimentales par assignation aléatoire réalistes, y compris les

recommandations pour de nouveaux protocoles itératifs qui pourraient être mis à jour en ligne aux différentes étapes de l'élaboration et de l'essai de la théorie de l'intervention.

Bibliographie

Bonell, Chris, Elizabeth Allen, Deborah Christie, Diana Elbourne et Adam Fletcher. 2014. « Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): study protocol for cluster randomised controlled trial ». *Trials* 15(381).

Bonell, Chris, Adam Fletcher, Natasha Fitzgerald-Yau, Daniel Hale, Elizabeth Allen et Diana Elbourne. 2015. « Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): pilot randomised controlled trial ». *Health Technol Assess* 19(53).

Bonell, Chris, Adam Fletcher, Matthew Morton, Theo Lorenc et Laurence Moore. 2012. « Realist randomised controlled trials: a new approach to evaluating complex public health interventions ». *Social Science & Medicine-Journals* 75(12) : 2299-2306.

Bonell, Chris, Adam Fletcher, Matthew Morton, Theo Lorenc et Laurence Moore. 2013. « Methods don't make assumptions, researchers do: a response to Marchal *et al.* ». *Social Science & Medicine-Journals* 94 : 81-82.

Byng, Richard, Ian Norman, Sally Redfern et Roger Jones. 2008. « Exposing the key functions of a complex intervention for shared care in mental health case study of a process evaluation ». *BMC Health Service Research* 8(274). doi : https://doi.org/10.1186/1472-6963-8-274.

Cartwright, Nancy, et Jeremy Hardie. 2012. *Evidence-based policy: a practical guide to doing it better*. Oxford: Oxford University Press.

Choksi, Dave A., et Thomas A. Farley. 2012. « The cost-effectiveness of environmental approaches to disease prevention ». *New England Journal of medicine* 367(4) : 295-97.

Craig, Peter, Paul Dieppe, Sally Macintyre, Susan Michie, Irwin Nazareth et Mark Petticrew. 2008. « Developing and evaluating complex interventions: the new Medical Research Council guidance ». *BMJ* 337(a1655). doi : https://doi.org/10.1136/bmj.a1655.

Hawe, Penelope, Alan Shiell et Therese Riley. 2009. « Theorising interventions as events in systems ». *American Journal of Community Psychology* 43(3-4) : 267-76. doi : https://doi.org/10.1007/s10464-009-9229-9.

Jamal, Farah, Adam Fletcher, Angela Harden, Helen Wells, James Thomas et Chris Bonell. 2013. « The school environment and student health: a meta-ethnography of qualitative research ». *BMC Public Health* 13(798).

Marchal, Bruno, Gill Westhorp, Geoff Wong, Sara Van Belle et Trisha Greenhalgh. 2013. « Realist RCTs of complex interventions – an oxymoron ». *Social Science & Medicine-Journals* 94 : 124-28. doi : https://doi.org/10.1016/j.socscimed.2013.06.025.

Moore, Graham F., Suzanne Audrey, Mary Barker, Lyndal Bond, Chris Bonell, et Wendy Hardeman. 2014. *Process evaluation of complex interventions: Medical Research Council guidance*. London: MRC Population Health Science Research Network.

Moore, Laurence, Suzanne Audrey, Mary Barker, Lyndal Bond, Chris Bonell et Wendy Hardeman. 2015. « Process evaluation of complex interventions: Medical Research Council guidance ». *BMJ* 350(h1258).

Pawson, Ray, et Nicholas Tilley. 1997. *Realistic evaluation*. London: Sage Publications.

Peto, Richard, Rory Collins et Richard Gray. 1995. « Large-scale randomised evidence: large and simple trials and overviews of trials ». *Journal of Clinical Epidemiology* 48 : 23-40.

6. Les essais randomisés réalistes peuvent-ils être authentiquement réalistes?

SARA VAN BELLE, GEOFF WONG, GILL WESTHORP, MARK PEARSON, NICK EMMEL, ANA MANZANO ET BRUNO MARCHAL

[Traduit de : Van Belle, Sara, Geoff Wong, Gill Westhorp, Mark Pearson, Nick Emmel, Ana Manzano & Bruno Marchal. 2016. « Can « realist » randomised controlled trials be genuinely realist? ». *Trials* 17 (313) : 1-6 (Extraits). Traduction par Carine Gazier et Valéry Ridde. Article originellement publié en *open access*.]

Jamal *et al.* (2015) proposent essentiellement des « méthodes expérimentales par assignation aléatoire réalistes » qui utilisent un processus technique (analyse statistique modératrice et médiatrice) pour tester des hypothèses sur les phénomènes observables. Toutefois, le réalisme part de l'hypothèse que tout ce qui a une importance significative ne peut être observé. La recherche réaliste a vocation à s'intéresser particulièrement aux processus causaux sous-jacents qui mènent aux résultats – ce que l'on appelle des mécanismes. Bien qu'ils ne soient pas directement observables, ni facilement mesurables, ils sont essentiels. D'un point de vue réaliste, ils sont nécessaires à l'explication du lien de causalité.

Qu'est-ce qu'un mécanisme?

La notion de mécanisme est essentielle pour comprendre l'étiologie et le traitement des maladies. Les mécanismes jouent également un rôle central dans la pensée réaliste. Ils y sont néanmoins conceptualisés différemment. Les interventions déclenchent des mécanismes dans des contextes spécifiques, ce qui aboutit à certains types de résultats. La nature des mécanismes a longtemps fait l'objet de discussions dans les milieux réalistes (Marchal *et al.*, 2012; Pawson, 1989), mais il existe un large consensus sur le fait que la « réaction aux ressources » est la caractéristique déterminante des mécanismes dans les travaux de Pawson et Tilley (Pawson et Tilley, 1997). D'un point de vue scientifique réaliste, les résultats de l'intervention peuvent être imputables aux ressources offertes et à la façon dont les intervenant-e-s et les participant-e-s y réagissent. En d'autres termes, faire et maintenir des choix différents nécessite un changement dans le raisonnement d'un-e participant-e (par exemple dans ses valeurs, ses croyances, ses attitudes ou la logique qu'il applique à une situation particulière) et/ou un changement dans les ressources (par exemple, informations, compétences, ressources matérielles ou soutiens) dont il ou elle dispose. Le réalisme affirme que c'est l'interaction entre le « raisonnement et les ressources » qui sous-tend les résultats de l'intervention (Greenhalgh *et al.*, 2015). En conséquence, les interventions fonctionnent de différentes manières pour différentes personnes car celles-ci réagissent différemment aux ressources fournies par l'intervention.

Du fait de leur nature, les mécanismes sont latents, invisibles et sensibles aux variations du contexte (Pawson, 2008). Les mécanismes peuvent se jouer au niveau des individus, des groupes, des organisations et de la société. On peut trouver des idées sur les mécanismes dans les théories psychologiques, sociales, culturelles, politiques et économiques (Astbury et Leeuw, 2010). Ainsi, le « mécanisme » au sens réaliste n'est pas assimilé

à des composantes de l'intervention, mais plutôt à la manière dont les ressources et les opportunités créées par l'intervention sont utilisées (ou non) par des personnes dans des contextes différents.

Jamal *et al.* (2015) définissent les mécanismes comme des aspects des interventions. Par exemple, à la page 2, Jamal *et al.* écrivent que « l'évaluateur ou l'évaluatrice doit poser des hypothèses et tester la manière dont la théorie de d'intervention du changement interagit avec le contexte pour permettre (ou empêcher) la mise en œuvre, les mécanismes causaux et, en fin de compte, les résultats ». Cependant, ce n'est pas la « théorie du changement » de d'intervention qui interagit avec le contexte. Le réalisme scientifique soutient plutôt que les interventions s'effectuent dans des contextes spécifiques et s'adressent aux acteurs et actrices qui décident (ou non) de modifier leur comportement, leurs choix ou leurs décisions en fonction des ressources et des opportunités offertes par l'intervention.

Les mécanismes sont également caractérisés par Jamal *et al.* (2015) comme des traitements externes. Les auteurs et autrices écrivent que « les évaluateurs et évaluatrices réalistes ont considéré que les interventions « fonctionnent » en introduisant des mécanismes qui interagissent avec les caractéristiques de leur contexte pour produire des résultats ». Cela implique, à tort, que les mécanismes peuvent être introduits dans une situation et donc être externes. Le réalisme scientifique soutient que les mécanismes ne sont pas des facteurs externes mais des aptitudes latentes, qui sont une fonction de l'interaction, entre les ressources d'intervention et les réponses des participant-e-s.

Tous les « mécanismes » présentés par Jamal et ses collègues (p. 8), sauf un, sont des activités ou la mise en œuvre d'actions. […] Cette confusion entre le concept de mécanisme et celui d'intervention (stratégie) est une erreur courante qui néglige les mécanismes réels en jeu (Marchal *et al.*, 2012; Pawson et Manzano, 2012).

Médiateurs et modérateurs ou configurations?

La conceptualisation du « mécanisme » est liée à l'approche analytique visant à identifier les mécanismes et à leur attribuer des effets. Dans plusieurs cas, Jamal *et al.* (2015) proposent d'utiliser des techniques d'analyse de la médiation pour identifier les mécanismes en jeu. Par exemple, ils et elles écrivent : « à ce stade, nous testerons les hypothèses dérivées des étapes 1 et 2 au moyen d'analyses quantitatives de la médiation des effets (pour examiner les mécanismes) et de la modération (pour examiner les conjonctures contextuelles) ». Ces auteurs et autrices écrivent également que « l'analyse de la médiation causale aide à identifier les variables de processus ou de médiation qui se trouvent dans les voies causales entre le traitement et le résultat… Les médiateurs sont des mesures post-intervention d'effets intermédiaires qui peuvent ou non tenir compte des effets de l'intervention sur les résultats finaux ».

Cette explication révèle la stratégie analytique des auteurs et autrices. Pour clarifier la manière dont leur approche contraste, voire s'oppose, à l'approche scientifique réaliste, nous nous tournons vers Mahoney (Mahoney, 2001). Cet auteur présente trois grandes façons de définir le « mécanisme » dans le domaine de la science. Premièrement, lorsque le terme « mécanisme causal » est utilisé dans les protocoles expérimentaux, il est généralement compris comme un (ensemble) de variable(s) intervenante(s) qui explique pourquoi il existe une corrélation entre une variable indépendante et une variable dépendante. Les mécanismes sont donc situés dans la boîte noire entre les variables indépendantes et dépendantes et s'expriment comme une variable. « Cependant, si la notion de mécanisme en tant que processus d'intervention est utile, cette définition ne va malheureusement pas au-delà des hypothèses de corrélation » (Mahoney, 2001) et, par conséquent, elle n'apporte que peu, ou pas, de preuves sur la causalité.

Mahoney identifie une deuxième définition du mécanisme qui « conçoit les mécanismes causaux comme des théories ou des variables de niveau intermédiaire qui peuvent être utilisées pour expliquer un éventail assez large de résultats » (Mahoney, 2001). Les mécanismes causaux ont été définis ici comme des « modèles causaux fréquents et facilement reconnaissables qui sont déclenchés dans des conditions généralement inconnues ou avec des conséquences indéterminées » (Elster, 1989), cité par Mahoney (2001). Cette définition se concentre sur les théories sous-jacentes du changement et ne propose pas de stratégie analytique spécifique pour démontrer l'effet du mécanisme. La première et la deuxième définition du mécanisme partagent une approche d'analyse corrélative de la causalité, qui met l'accent sur l'identification d'antécédents régulièrement associés aux résultats (mode d'explication causal successif).

Une troisième définition est utilisée par le réalisme scientifique. Elle considère un mécanisme causal comme « une entité non observée qui, une fois activée, génère un résultat d'intérêt » (Mahoney, 2001 : 581). C'est la conception de la causalité génératrice des mécanismes adoptés par la recherche réaliste, qui considère que les mécanismes sont des propriétés inhérentes à l'agencéité et aux structures.

> Faire et maintenir des choix différents exige une modification du raisonnement d'un participant (par exemple dans ses valeurs, ses croyances, ses attitudes ou la logique qu'il applique à une situation particulière) et/ou des ressources (par exemple Informations, compétences, ressources matérielles, soutiens) dont il dispose. Cette combinaison de raisonnement et de ressources est ce qui permet au programme de « fonctionner » et est connu sous le nom de « mécanisme » (Greenhalgh *et al.*, 2015).

Cette conceptualisation génératrice fait passer l'analyse de la causalité au-delà de l'analyse corrélative.

Jamal *et al.* (2015) semblent avoir adopté la première définition qui réduit les mécanismes (et aussi le « contexte ») à de simples variables – bien qu'ils et elles se dirigent peut-être vers la deuxième définition. En tout état de cause, leur utilisation des termes « médiation » et « modération » implique une approche de l'analyse orientée vers les variables, contrairement à la perspective orientée vers la configuration qui reconnaît la causalité complexe adoptée dans le réalisme scientifique. Jamal *et al.* (2015) étendent l'approche axée sur les variables à la formulation des hypothèses. Les auteurs et autrices présentent la façon dont ils et elles ont développé un ensemble de mécanismes d'intervention, dits pré-hypothèses (qu'ils et elles qualifient également d'hypothèses de médiation), séparément d'un ensemble d'obstacles et médiateurs contextuels. Nous ne pouvons que supposer que les auteurs et autrices ont suivi cette voie, car il serait plus facile de tester statistiquement les différentes hypothèses en tant que volets distincts. Cependant, le développement de configurations contexte-mécanisme-résultat va au-delà de la « segmentation » de la théorie du programme en une série de variables sur le contexte et une autre sur le traitement de ce qu'ils appellent les « mécanismes d'intervention ». Dans le réalisme scientifique, l'explication repose sur la démonstration de la relation entre le contexte et le mécanisme. Les hypothèses devraient présenter un ensemble de théories de programme qui expliquent comment les modèles de résultats peuvent être expliqués par une configuration d'interventions, d'acteurs, de contextes et de mécanismes. Cela reflète la reconnaissance d'une causalité complexe dans l'évaluation réaliste.

Cela ne signifie pas que la recherche réaliste s'oppose à l'utilisation de données quantitatives. La recherche réaliste ne conçoit pas les mesures quantitatives comme des « variables » (dans le sens de choses qui varient en quantité et provoquent une variation ultérieure dans l'élément suivant de l'équation) mais comme des « indicateurs » (dans le sens d'une mesure partielle d'un aspect d'une chose qui « indique » si elle est présente et/ou la mesure dans laquelle elle est présente). En utilisant cette dernière définition, il est tout à fait possible d'utiliser des mesures quantitatives

pour n'importe quel contexte (C), mécanisme (M) ou résultat (R), en supposant bien sûr que le C, le M et le R ont été théorisés antérieurement et que les indicateurs utilisés sont « adaptés à l'objectif » de cette théorie. Il s'agit d'une utilisation configuratrice des indicateurs plutôt que d'une analyse basée sur des variables.

En résumé, dans le réalisme scientifique, l'analyse ne dépend pas de l'évaluation statistique de la corrélation entre les variables représentant l'intervention, l'effet, les modérateurs et les médiateurs. L'analyse utilise plutôt toutes les données et les méthodes d'analyse appropriées pour élaborer, soutenir, refaire ou affiner des explications plausibles qui intègrent l'intervention, les acteurs et actrices, les résultats, le contexte et les mécanismes. Cela nous amène à notre troisième façon de considérer le réalisme scientifique, distincte de la méthodologie proposée par Jamal et ses collègues.

Les méthodes expérimentales par assignation aléatoire peuvent-elles tenir compte des configurations dynamiques de la CMR?

La dernière différence avec l'article de Jamal *et al.* (2015) concerne la capacité de la méthode expérimentale par assignation aléatoire à donner un sens à l'interaction dynamique entre l'intervention, les acteurs et actrices, le contexte et les mécanismes qui, d'un point de vue réaliste, contribuent à certains types de résultats au sein d'une intervention complexe. Si l'on considère que les configurations de la CMR (comme proposé par Pawson et Tilley) peuvent être évaluées dans une méthode expérimentale par assignation aléatoire, alors ce protocole de recherche devrait être en mesure de démontrer comment et pourquoi, certains types de résultats sont causés par des mécanismes « déclenchés » dans des contextes spécifiques. Là encore, les mécanismes doivent être compris dans une perspective réaliste et non pas comme une chaîne

de facteurs entre l'intervention et le résultat. Le protocole de recherche devrait également être en mesure de démontrer comment et pourquoi ces configurations de la CMR varient selon les personnes ou les contextes et évoluent dans le temps. Compte tenu de la nécessité d'une assignation aléatoire et d'un contrôle dans la méthode expérimentale par assignation aléatoire, seules des configurations de la CMR simples et relativement peu nombreuses peuvent être testées simultanément. Dans le meilleur des cas, la méthode expérimentale par assignation aléatoire pourrait nous aider à évaluer la contribution relative des mécanismes à certains types de résultats si la configuration causale est uniforme, mais pas lorsqu'il est probable que des mécanismes différents généreront des résultats différents dans des circonstances différentes, comme c'est la règle plutôt que l'exception dans toute intervention sanitaire.

Bibliographie

Astbury, Brad, et Frans L. Leeuw. 2010. « Unpacking black boxes: mechanisms and theory building in evaluation ». *American Journal of Evaluation* 31(3) : 363-81. doi : https://doi.org/10.1177%2F1098214010371972.

Elster, Jon. 1989. *Nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.

Greenhalgh, Trisha, Geoff Wong, Joanne Greenhalgh, Justin Jagosh, Ana Manzano et Ray Pawson. 2015. « Protocol—the RAMESES II study: developing guidance and reporting standards for realist evaluation ». *BMJ Open*.

Jamal, Farah, Adam Fletcher, Nichola Shackleton, Diana Elbourne, Russell Viner et Chris Bonell. 2015. « The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example ». *Trials* 16(466). doi : https://doi.org/10.1186/s13063-015-0980-y.

Mahoney, Joseph L. 2001. « Beyond correlational analysis: recent innovations in theory and method ». *Sociological Forum* 16(3) : 575-93.

Marchal, Bruno, Sara Van Belle, Josefien Van Olmen, Tom Hoerée et Guy Kegels. 2012. « Is realist evaluation keeping its promise? A literature review of methodological practice in health systems research ». *Evaluation* 18(2) : 192-212. doi : https://doi.org/10.1177%2F1356389012442444.

Pawson, Ray. 1989. A *measure for measures*: *a manifesto for empirical sociology*. London: Routledge.

Pawson, Ray. 2008. « Invisible mechanisms ». *Australasian Evaluation Society* 8(2).

Pawson, Ray, et Ana Manzano. 2012. « A realist diagnostic workshop ». *Evaluation* 18(2) : 176-91.

Pawson, Ray, et Nicholas Tilley. 1997. *Realistic evaluation*. London: Sage Publications.

7. Sur les essais réalistes et la mise à l'épreuve des configurations contexte – mécanismes – résultats : en réponse à Van Belle et al.

CHRIS BONELL, EMILY WARREN, ADAM FLETCHER ET RUSSELL VINER

[Traduit de : Bonell, Chris, Emily Warren, Adam Fletcher, Russell Viner. 2016. « Realist trials and the testing of context-mechanism-outcome configurations: a response to Van Belle *et al.* ». *Trials* 17 (478) : 1-5 (Extraits). Traduction par Carine Gazier et Valéry Ridde. Article originellement publié en *open access*.]

Une compréhension réaliste des interventions et des mécanismes

Soyons tout d'abord clairs sur la nature de l'intervention « apprendre ensemble » dont il est question dans notre article précédent (Jamal *et al.* 2015) et sur la manière dont elle est censée fonctionner. Nous sommes tout à fait d'accord avec Van Belle *et al.* (2016) pour dire que les interventions comprennent une série de ressources. « Apprendre ensemble » vise à réduire le harcèlement et les violences dans les établissements d'enseignement secondaire en mettant à leur disposition les ressources suivantes : (1) des plans de cours et des diapositives pour un programme de compétences sociales et émotionnelles; (2) un dossier sur les besoins locaux des élèves, un manuel et un médiateur ou une médiatrice externe; et (3) des séances de formation sur les pratiques réparatrices à l'intention du personnel. Nous offrons ces ressources aux établissements dans l'espoir que le personnel et les élèves décident de

les utiliser afin de faciliter diverses actions telles que : des cours sur les compétences sociales/émotionnelles; des réformes des politiques scolaires et d'autres décisions prises localement à l'échelle de l'établissement; et des séances sur les pratiques réparatrices au cours desquelles le personnel et les élèves réagissent à partir de faits particuliers de harcèlement ou de violences. Nous faisons l'hypothèse qu'à travers ces actions divers mécanismes peuvent être déclenchés. Nous sommes d'accord avec Van Belle *et al.* (2016) pour dire que les interventions sociales ne peuvent pas imposer directement des mécanismes; ceux-ci ne peuvent s'enclencher que par le biais d'un processus par lequel les participant-e-s agissent sur les ressources fournies. Dans le cas d' « apprendre ensemble », il existe divers mécanismes qui impliquent l'érosion de diverses « frontières » entre et parmi le personnel et les élèves, mais aussi entre le développement scolaire des élèves et le développement en général. Nous faisons l'hypothèse que ces frontières seront érodées non pas directement par l'intervention, mais par, d'une part, l'engagement du personnel et des élèves dans des actions soutenues par les ressources de l'intervention, pour améliorer les relations dans l'établissement et par, d'autre part, la réorientation des activités scolaires de manière à mettre l'accent sur le développement holistique des élèves, qui devrait inclure les résultats scolaires s'en s'y réduire.

Sur la base de la théorie du fonctionnement humain et de l'organisation scolaire Markham et Aveyard (2003) ainsi que des recherches qualitatives sur l'environnement scolaire et la santé des jeunes (Jamal *et al.* 2013), nous avons émis l'hypothèse, dans notre article original, que l'érosion de ces frontières encouragerait davantage d'élèves, en particulier ceux issu-e-s des classes populaires, à se sentir attaché-e-s à l'éducation, moins attaché-e-s à des groupes de pairs anti-scolaires et moins engagé-e-s dans des pratiques allant à l'encontre des règles scolaires formelles et des normes informelles, y compris le harcèlement et les violences (Jamal *et al.* 2015). Nous avons en outre pensé que ces mécanismes seraient déclenchés et fonctionneraient différemment dans différents contextes.

Par exemple, nous avons émis l'hypothèse que dans les établissements où le personnel accorde déjà une certaine priorité à la promotion du bien-être général des élèves, les activités d'intervention seront plus susceptibles d'être mises en œuvre, de sorte que les mécanismes d'intervention, notamment ceux concernant l'érosion des frontières entre le développement scolaire et le développement général des élèves, seront plus susceptibles d'être déclenchés. Enfin, nous avons émis l'hypothèse que dans les établissements où les élèves d'origine modeste sont plus nombreux, l'érosion des frontières entrainera une plus grande proportion d'élèves à s'impliquer dans leur éducation (parce que nous supposons que les frontières susmentionnées entravent en particulier l'engagement scolaire des élèves issus de milieux sociaux défavorisés). Nous espérons que cette description plus complète de notre intervention et de ses mécanismes rassurera les lecteurs et lectrices sur le fait que notre compréhension de l'intervention « apprendre ensemble » est compatible avec une ontologie et une pratique d'évaluation réalistes.

Une compréhension réaliste de l'évaluation empirique

Nous continuons d'affirmer que notre approche de la recherche en général et de l'évaluation en particulier est réaliste dans son orientation. Les réalistes suggèrent qu'il existe un domaine « empirique » réel composé des données que les chercheurs collectent et analysent. Ce domaine empirique réel fournit une fenêtre, bien qu'indirecte, sur un domaine du « réel » d'événements apparents pour les participants. Ce domaine du réel reflète à son tour un domaine « réel » constitué de mécanismes structurels non observables, mais qui sont les causes des domaines réels et empiriques (Bhaskar, 1989). S'agissant d'épistémologie, les réalistes croient qu'ils et elles peuvent identifier des vérités objectives décrivant le domaine du réel et découvrir les véritables mécanismes causaux qui en relèvent en se basant sur des données provenant du domaine empirique. Comme Van Belle *et al.* (2016), nous pensons que

même si des mécanismes d'intervention se produisent, ils ne seront pas directement observables. Les « frontières » que l'intervention cherche à éroder, de même que l'engagement scolaire, sont dans le domaine du réel. Ils provoquent des phénomènes observables, mais ne sont pas eux-mêmes observables.

Les activités que notre intervention vise à promouvoir sont, dans un langage réaliste critique, dans le domaine du réel, de même que les résultats que nous espérons obtenir à la suite de l'intervention. Le personnel et les élèves pourront observer les activités d'intervention, telles que les séances sur les pratiques réparatrices, ainsi que les comportements, tels que le harcèlement, que l'intervention vise à réduire. Cependant, les données recueillies dans le cadre de la méthode expérimentale par assignation aléatoire d'« apprendre ensemble » (comme les données recueillies dans toutes formes de recherche) ne constituent pas une fenêtre directe et sans problème sur ce domaine du réel. Notre évaluation des résultats ne consiste pas à recueillir directement des données sur le harcèlement ou les agressions. Il s'agit plutôt de recueillir des données sur les réponses des élèves aux questionnaires leur demandant leur expérience de ces pratiques. De même, notre évaluation des processus ne permet pas de recueillir directement des données sur des activités telles que la révision par l'établissement de ses politiques et les sessions sur les pratiques réparatrices. Elle recueille plutôt des données sous la forme de notes que les chercheurs et chercheuses prennent lorsqu'ils et elles observent ces séances ou sous la forme de comptes rendus d'enseignants ou d'élèves lorsqu'ils et elles sont interrogé-e-s sur leur expérience de ces activités. S'agissant du réalisme critique, ces données sont dans le domaine empirique. Nous comprenons que toutes sortes de facteurs peuvent signifier que ces données ne fournissent pas une représentation complète ou sans problème des événements dans le domaine du réel. Néanmoins, elles devraient fournir des indications sur ce qui se passe.

Sur la base de notre théorie concernant la façon dont les mécanismes (dans le domaine du réel) interagissent avec le contexte pour produire des résultats (dans le domaine du réel), nous avons émis l'hypothèse que dans les analyses statistiques des indicateurs de résultats (dans le domaine empirique), les élèves des établissements choisis au hasard pour mettre en œuvre l'intervention déclareront moins de harcèlement et de violences que les élèves des établissements choisis au hasard pour être des témoins (Jamal *et al.*, 2015). Nous avons également supposé que dans les analyses de médiation, l'association entre le groupe expérimental et les indicateurs empiriques du harcèlement et des violences serait réduite par l'ajustement des indicateurs de l'engagement scolaire accru des élèves. En outre, nous avons émis l'hypothèse que dans les analyses de modération statistique, les indicateurs de base agrégés au niveau de l'établissement concernant les priorités du personnel, et les indicateurs au niveau de l'établissement concernant le statut socioéconomique des élèves, modéreront l'association constatée entre le volet d'intervention et nos mesures contre le harcèlement et les violences.

Les exemples ci-dessus illustrent la façon dont nous entendons utiliser des analyses statistiques (dans le domaine empirique) pour tester des hypothèses sur la façon dont, dans le domaine du réel, le contexte et les mécanismes interagissent pour générer des résultats (connus sous le nom de contexte-mécanisme-résultats ou « configurations CMR »). Van Belle *et al.* (2016) ne disent pas grand-chose sur les raisons pour lesquelles il est peu probable que les essais randomisés réalistes soient en mesure de tester empiriquement les hypothèses relatives à la CMR. Ils et elles affirment, sans se référer à des preuves ou à d'autres arguments, qu'« étant donné la nécessité d'une assignation aléatoire et d'un contrôle dans une méthode expérimentale par assignation aléatoire, seules des configurations de CMR simples et relativement peu nombreuses peuvent être testées simultanément ». Nous ne sommes pas d'accord. Nous avons l'intention d'entreprendre des analyses pour tester plusieurs CMR, y compris, mais non exclusivement, celles mentionnées ci-dessus. Certaines d'entre elles reposent *a priori* sur la théorie, tandis que d'autres

ont été et continueront d'être développées sur la base de recherches qualitatives dans le cadre d'essais. Il devrait être parfaitement possible de les tester dans le cadre d'une méthode expérimentale par assignation aléatoire. Ces analyses devraient nous aider à mieux comprendre comment et pourquoi la réduction du harcèlement et des violences est due à des mécanismes d'érosion des frontières et d'engagement scolaire qui sont déclenchés et se déroulent différemment dans divers contextes scolaires. Toutefois, la question de savoir si c'est le cas ou non est en fin de compte une question empirique. Nous ne pourrons pas en juger tant que nous n'aurons pas terminé nos analyses. Nous ne voyons pas, dans la répartition aléatoire, ce qui pourrait empêcher de telles analyses. En effet, faire ces analyses dans le cadre d'une méthode expérimentale par assignation aléatoire présente l'avantage crucial de contrôler les facteurs confondants. C'est très important en santé publique, car même les interventions importantes ont souvent des effets très limités pour chaque participant-e, qu'il est impossible de distinguer des autres influences (Cousens *et al.*, 2011). Nous convenons que notre capacité à évaluer les configurations de CMR serait compromise si les essais ne contenaient pas suffisamment de variété dans les caractéristiques de lieux ou de personnes en raison de critères d'inclusion trop stricts pour les groupes d'échantillonnage ou les individus. Mais si cela peut parfois se produire dans le cadre d'essais, ce n'est pas une caractéristique nécessaire, en particulier dans le cas d'essais d'efficacité pragmatique. Les autres obstacles évidents aux analyses proposées sont l'erreur de mesure et le manque de capacité statistique. Il s'agit là de véritables défis, mais ils ne sont en aucun cas une caractéristique particulière des essais, contrairement à d'autres protocoles de recherche.

De plus, nous tenons à souligner que l'élaboration et la vérification empirique de telles hypothèses ne signifient pas, comme l'indiquent Van Belle *et al.* (2016), que nous confondons l'empirique avec le réel, en réduisant le mécanisme causal de notre intervention aux analyses de médiation statistique ou en réduisant le contexte de l'intervention aux analyses de modération statistique. Nous utilisons des données

quantitatives brutes et indirectes (qui existent dans le domaine empirique) pour vérifier indirectement si nos théories sur les mécanismes (qui existent dans le domaine du réel) peuvent être correctes. En outre, nous utilisons également des recherches qualitatives pour approfondir notre compréhension du fonctionnement des mécanismes et pour affiner nos théories et nos hypothèses. À l'instar de nos recherches quantitatives, nos recherches qualitatives examinent les données empiriques (c'est-à-dire les comptes rendus des étudiants et du personnel) comme une fenêtre indirecte sur le concret (expériences des participant-e-s) et le réel (la façon dont les mécanismes se déploient à travers les interactions entre les structures individuelles et les structures sociales).

En outre, si nos analyses ne corroborent pas les hypothèses ci-dessus, cela ne signifie pas que nous concluons immédiatement que les mécanismes théorisés n'existent pas. Des résultats nuls pourraient indiquer que le contexte fait que les mécanismes ne sont pas activés (par exemple, les établissements ne mettent pas en œuvre l'intervention ou la mise en œuvre ne déclenche pas une érosion des frontières) ou que le mécanisme est activé, mais contrecarré par d'autres (par exemple, les initiatives gouvernementales poussent les établissements à renforcer les frontières entre le développement scolaire et le développement général des élèves). Il convient de souligner que cette approche est conforme à la philosophie actuelle sur la façon d'interpréter les résultats nuls des méthodes expérimentales par assignation aléatoire des interventions sociales (Cartwright et Hardie, 2012).

Bibliographie

Bhaskar, Roy. 1989. *The Possibility of Naturalism: A Philosophical Critique of the Contemporary Human Sciences*. 2e éd. Hemel Hempstead: Harvester Wheatsheaf.

Cartwright, Nancy, et Jeremy Hardie. 2012. *Evidence-based policy: a practical guide to doing it better*. Oxford: Oxford University Press.

Cousens, Simon, James R. Hargreaves, Chris Bonell, J. Thomas, Betty Kirkwood et Richard J. Hayes. 2011. « Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference ». *Journal of Epidemiology and Community Health* 65 : 576-81. doi : https://doi.org/10.1136/jech.2008.082610.

Jamal, Farah, Adam Fletcher, Angela Harden, Helen Wells, James Thomas et Chris Bonell. 2013. « The school environment and student health: a meta-ethnography of qualitative research ». *BMC Public Health* 13(798).

Jamal, Farah, Adam Fletcher, Nichola Shackleton, Diana Elbourne, Russell Viner et Chris Bonell. 2015. « The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example ». *Trials* 16(466). doi : https://doi.org/10.1186/s13063-015-0980-y.

Markham, Wolfgang A., et Paul Aveyard. 2003. « A new theory of health promoting schools based on human functioning, school organisation and pedagogic practice ». *Social Science & Medicine – Journals* 56 : 1209-20.

Van Belle, Sara, Geoff Wong, Gill Westhorp, Mark Pearson, Nick Emmel, Ana Manzano et Bruno Marchal. 2016. « Can “realist” randomised controlled trials be genuinely realist? » *Trials* 17(313).

Le regard de Manuela De Allegri

MANUELA DE ALLEGRI

Lorsqu'en 2019, le prix Nobel en économie a été attribué à Abhijit Banerjee, Esther Duflo et Michael Kremer, j'ai ressenti des émotions contrastées. D'une part, en tant qu'économiste ou du moins économiste partielle, et tout aussi engagée dans la réduction de la pauvreté, j'étais heureuse à l'idée qu'un prix aussi prestigieux ait été attribué à des personnes engagées dans la production de preuves sur ce qui fonctionne et ce qui ne fonctionne pas dans la lutte contre la pauvreté. D'un autre côté, je craignais que la déclaration accompagnant le prix, « pour leur approche expérimentale de la réduction de la pauvreté dans le monde », ne renforce l'idée que les méthodes expérimentales et quasi-expérimentales représentent le seul mode de production de preuves crédibles lors de l'évaluation des interventions sociales. L'hypothèse implicite de cette déclaration est que la complexité du monde dans lequel nous vivons peut être réduite à une série de questions de recherche étroitement définies auxquelles il faut répondre par des expériences soigneusement conçues. En d'autres termes, la déclaration et le prix qui l'accompagnait ont renforcé le message selon lequel une série d'essais contrôlés randomisés bien planifiés serait tout ce dont nous avons besoin pour donner un sens à notre réalité et trouver des solutions à nos plus grands problèmes.

En réfléchissant à cette déclaration et à ses implications, j'ai compris qu'il était temps pour moi de repenser au type de chercheuse que je voulais être et, plus important encore, au type d'exemple que je voulais donner en tant que mentor des jeunes générations de chercheur-e-s engagé-e-s dans l'évaluation. Cette déclaration m'a ramené au jour, bien des années auparavant, où, en tant que jeune post-doc, j'ai vu mon superviseur de l'époque dire à la doctorante engagée sur le même projet que moi que l'analyse qualitative sur laquelle ils avaient travaillé était d'une importance

secondaire pour notre travail, que ce qui comptait vraiment était l'évaluation économique d'impact que je menais. Je reconnais que cette rencontre a marqué le moment où j'ai pris conscience de la manière dont notre positionnement vis-à-vis des méthodes de recherche que nous employons dans notre travail révèle nos convictions épistémologiques les plus profondes sur ce qui constitue une recherche valide et sur le rôle que nous assumons en tant que chercheuses et chercheurs.

Ayant été formée, tant au niveau du premier cycle que du troisième cycle, par des chercheurs et chercheuses dédié-e-s aux méthodes mixtes qui avaient adopté le pragmatisme comme principal paradigme bien avant qu'il ne devienne aussi populaire qu'aujourd'hui, j'avais été protégée d'une exposition directe à des positionnements épistémologiques plus tranchés. J'avais lu des articles sur le positivisme et le constructivisme, et j'avais pris connaissance de certaines communautés de recherche qui revendiquaient la suprématie de la méthode quantitative ou qualitative sur l'autre. Mais l'expérience directe de quelque chose est très différente de la lecture d'un texte. Ce que cette rencontre m'a révélé, et ce que le prix Nobel d'économie m'a rappelé dix ans plus tard, c'est que nous ne parlons pas ouvertement de nos convictions épistémologiques, mais qu'elles dominent la manière dont nous abordons la production de connaissances et les méthodes que nous choisissons d'employer pour réaliser une évaluation. Nos convictions épistémologiques profondément ancrées déterminent les preuves que nous produisons et que nous intégrons dans la formulation des politiques sociales, avec un impact profond sur la vie de millions de personnes. Nous ne pouvons pas considérer les convictions épistémologiques et les options méthodologiques qui en découlent comme une question d'importance mineure, pertinente uniquement dans le cadre d'un discours académique, car elles imprègnent les preuves que nous produisons et qui servent comme base à la prise de décision politique.

Au fil des ans, j'ai découvert que les convictions épistémologiques que nous avons sont étroitement liées à la formation que nous avons reçue. L'introduction de la partie Paradigmes et la sélection d'articles traduits

qui l'accompagne nous invitent à réfléchir à la façon dont, selon la tradition académique dans laquelle nous avons été formés, nous conceptualisons différemment la causalité et abordons donc l'évaluation avec des perspectives différentes, en employant des conceptions différentes, des outils de collecte de données différents et des stratégies analytiques différentes. Tout d'abord, en simplifiant volontairement la complexité des écrits, la partie Paradigmes nous amène à nous demander si notre propre tradition scientifique reconnaît ou conteste le concept de « contrefactuel » comme principe central de la recherche en évaluation et, par conséquent, si elle nous encourage à utiliser des méthodes expérimentales ou plutôt réalistes dans notre pratique de l'évaluation. Après une synthèse réfléchie des points centraux de ce débat, la partie nous propose de sortir des restrictions imposées par notre propre tradition scientifique, en nous suggérant de faire un saut dans l'adoption de solutions plus innovantes, qui nous permettent d'intégrer des traditions épistémologiques apparemment contradictoires dans un paradigme de recherche unique. Ces approches, telles que les essais contrôlés randomisés réalistes, nous permettent de concilier la logique positiviste de l'essai contrôlé randomisé avec celle de la logique réaliste visant à explorer et à expliquer les mécanismes du changement. La partie examine ensuite comment ces approches visent à libérer la pratique de l'évaluation de tout dogmatisme, en adoptant une approche pragmatique pour répondre aux questions pertinentes en répondant à la fois à la nécessité d'établir des relations de cause à effet et à celle d'expliquer les mécanismes par lesquels le changement s'est produit.

En tant que personne ayant construit sa carrière en travaillant à la croisée des paradigmes et des approches méthodologiques qui en découlent, en plaidant pour le pragmatisme comme seul guide pour définir les modèles et les méthodes à appliquer pour décider de la meilleure façon de répondre à une question de recherche donnée, je ne pourrais pas être plus d'accord avec la perspective proposée à la fin de l'introduction. Je vois dans l'oxymore consistant à concilier les approches expérimentales et le réalisme critique une occasion unique de faire avancer le discours

sur l'évaluation de manière significative dans le sens où elle sert l'objectif de générer des preuves pertinentes pour les politiques plutôt que d'être au service d'une communauté de recherche spécifique et de ses intérêts. Je souhaite moi-même un monde où la formation à l'évaluation serait dissociée de toute tradition académique spécifique, afin que les personnes engagées dans une évaluation puissent commencer par formuler librement les questions de recherche pertinentes, sans craindre de ne pas pouvoir y répondre, parce que les méthodes nécessaires dépasseraient les prescriptions dogmatiques de leur domaine disciplinaire spécifique. Au fil du temps, j'en suis venue à apprécier de plus en plus la liberté que procure l'accès à une vaste boîte à outils, en choisissant les approches et les méthodes en fonction des besoins spécifiques de l'évaluation à laquelle je suis confrontée. J'aime mélanger les approches de manière peu orthodoxe et encourager mes étudiant-e-s à faire de même, convaincue que ce n'est qu'en agissant de la sorte que nous pouvons véritablement repousser les limites de la connaissance et générer des évaluations informatives.

Il faut toutefois préciser qu'être pragmatique ne signifie pas, à mon avis, ignorer les traditions épistémologiques à la base des choix méthodologiques que nous faisons. Je constate que de plus en plus de personnes procèdent à des évaluations sans avoir une idée claire des positions épistémologiques qui sont à la base des différentes méthodes et ne parviennent donc pas à saisir la logique de leur application. De telles pratiques me préoccupent beaucoup, car je crains que, sans une compréhension claire des fondements conceptuels de notre travail, nous ne puissions pas définir des modèles, des outils de collecte de données et des stratégies analytiques qui aboutissent à des évaluations informatives et significatives. Par conséquent, si je reconnais la nécessité de nous libérer des prescriptions dogmatiques et de nous permettre de mélanger les approches comme la seule façon d'avancer dans l'évaluation, je reconnais également la nécessité d'être prudent-e et attentive ou attentif lorsque nous nous engageons dans ce processus. J'aime à penser qu'il s'agit là de la vigilance du chercheur ou de la chercheuse, conscient-e des

paradigmes à l'origine des méthodes à notre disposition pour nous libérer de leurs prescriptions dogmatiques, tout en restant fidèle à la logique de leur application. Lorsque je considère le défi de la mise en œuvre d'une telle approche dans la pratique, je reviens souvent au livre de Jennifer C. Greene (Greene, 2007), qui suggère que la voie à suivre dans la pratique de l'évaluation réside dans notre capacité à laisser de multiples modèles mentaux entrer dans un seul exercice d'évaluation.

Dans la lignée de cette pensée, j'aimerais encourager les étudiant-e-s en évaluation à se familiariser avec les constructions existantes et à accueillir toutes ces constructions pour qu'elles cohabitent dans leur esprit. En pratique, cela signifie inviter le positiviste qui apprécie la clarté des expériences, le constructiviste social qui aime explorer la complexité et les réalités multiples, et le réaliste critique qui cherche à comprendre les mécanismes d'action qui cohabitent chez une même personne. En les regardant débattre dans votre tête sur l'approche à adopter pour une évaluation donnée, gardez votre attention exclusivement sur la question de recherche reflétant le problème à résoudre. C'est de ce débat fructueux que naîtra la voie à suivre.

Bibliographie

Greene, Jennifer C. 2007. *Mixed Methods in Social Inquiry*. San Francisco: Jossey-Bass.

Liste des auteurs et autrices

Marvin C. Alkin est professeur émérite en éducation à l'Université de Californie à Los Angeles (UCLA). Théoricien de l'évaluation, il a notamment travaillé sur les questions d'utilisation de l'évaluation, et plus largement sur la classification des théories et théoricien-ne-s de l'évaluation. Son nom est en particulier associé à l'arbre des théories d'évaluation, dont il a signé la première version avec Christina Christie.

Thomas Archibald est professeur agrégé, spécialiste de la vulgarisation et directeur du programme d'études supérieures au *Agricultural, Leadership, and Community Education Department* à Virginia Tech, où il dirige également le projet Feed the Future Senegal Jeunesse en Agriculture. Ses recherches et sa pratique se concentrent principalement sur le renforcement des capacités d'évaluation, la pensée évaluative et le développement positif des jeunes. Lauréat du prix « *Marcia Guttentag Promising New Evaluator Award* » de l'Association American de l'Evaluation (AEA) et du Virginia Tech « *Alumni Award for Excellence in International Outreach* », il est membre du conseil d'administration de l'*Eastern Evaluation Research Society* et est rédacteur en chef adjoint de la revue Evaluation and Program Planning. Il a obtenu son doctorat en éducation des adultes et de vulgarisation de l'Université Cornell en 2013.

Michael Bamberger a travaillé dans des ONG puis à la Banque mondiale avant de devenir consultant indépendant. Ses apports portent en particulier sur le développement d'approches adaptées aux conditions réelles d'exercice de l'évaluation d'impact dans le monde du développement, intégrant méthodes quantitatives et qualitatives.

Chris Bonnell est professeur en santé publique et en sociologie à la *London school of hygiene and tropical medicine*. Spécialisé en méthodologie de l'évaluation, ses recherches portent notamment sur la santé des adolescent-e-s et la santé sexuelle.

Donald T. Campbell (1916-1996), professeur en psychologie, est un fervent promoteur d'une société expérimentale, dans lesquelles les innovations sociales pourraient être testées, et retenues ou non au regard des résultats de l'évaluation. Il est connu pour son rôle, avec Thomas D. Cook, dans l'importation des approches expérimentales et quasi-expérimentales dans le champ de l'évaluation. Il est à l'origine des concepts de validité interne et externe et s'intéresse en général aux biais qui affectent les résultats d'évaluation.

Mary Church est docteure en psychologie expérimentale. Après une expérience de 10 ans en évaluation de programme en éducation et santé publique dans le Pacifique, elle est désormais psychologue clinicienne et exerce en pratique privée à Hawaï.

Thomas D. Cook est professeur émérite de sociologie à l'Université Northwestern. Il a été un de ceux, avec Donald Campbell, à avoir importé dans le champ de l'évaluation les méthodes d'inférence causale expérimentales et quasi-expérimentales. Il a en particulier été actif dans le champ de l'évaluation des programmes éducatifs. Il a eu un rôle important pour consolider des éléments de théorie de l'évaluation, notamment avec William Shadish.

Gary B. Cox est professeur au département de psychiatrie de l'Université de Washington.

Manuela De Allegri a une formation universitaire en sociologie, économie de la santé et santé publique. Elle est responsable du groupe de recherche en économie de la santé et financement de la santé à l'Institut en Santé Globale de Heidelberg, Allemagne. Ses domaines d'expertise comprennent le financement de la santé, l'évaluation de l'impact et des processus, et l'évaluation économique.

Thomas Delahais est évaluateur et cofondateur de la société coopérative Quadrant Conseil. Ses travaux portent sur l'évaluation des interventions complexes et en particulier sur l'analyse de contribution, sur l'évaluation des initiatives de transition et sur la sociologie de l'évaluation. Il est membre du bureau éditorial du *Evaluation Journal*.

Agathe Devaux-Spatarakis est consultante et chercheuse pour la Scop Quadrant Conseil. Agathe conduit des missions d'évaluation de politiques publiques et d'accompagnement méthodologique pour le compte d'organisations publiques en France et à l'international. Docteure en science politique, elle est spécialisée dans les méthodes d'évaluation adaptées aux innovations et expérimentations ainsi que l'étude de l'utilisation des résultats par les décideurs publics.

Diana Elbourne est professeure en évaluation des soins de santé à la *London school of hygiene and tropical medicine*, spécialisée dans les essais randomisés contrôlés.

Nick Emmel est professeur en méthodologie des sciences sociales à l'Université de Leeds. Il est un des leaders de l'approche réaliste en évaluation.

Adam Fletcher est enseignant à l'Université de Cardiff et directeur adjoint du centre d'excellence en santé publique DECIPHer. Il est spécialiste de l'évaluation réaliste.

Lucia Fort est consultante à la Banque mondiale, spécialisée dans les questions de genre et de développement social.

Yves Gingras est professeur en histoire et en sociologie à l'Université du Québec à Montréal (UQAM). Il a cofondé en 1997 l'Observatoire des sciences et des technologies, dont il assure la direction scientifique.

Jennifer C. Greene est professeure de psychologie de l'éducation à l'Université de l'Illinois. Elle est une théoricienne et une praticienne de l'évaluation. Dans ses travaux, elle a notamment visé à légitimer les

approches participatives et à donner corps à une meilleure prise en compte des valeurs des parties prenantes dans l'évaluation, qui se doit de se faire la voix des plus marginalisés. Elle a aussi largement travaillé sur l'usage des méthodes mixtes en évaluation.

Egon G. Guba (1924-2008) a très tôt remis en cause les cadres classiques de l'investigation évaluative. Avec Yvonna Lincoln, il se fait le théoricien d'une évaluation naturaliste, d'inspiration constructiviste, dans laquelle les problèmes soulignés par les parties prenantes sont étudiés, de façon itérative, avec pour objectif d'aboutir à des consensus d'interprétation.

Gary T. Henry est professeur en éducation et doyen de la faculté d'éducation de l'université du Delaware. Il a mené de nombreuses évaluations dans le champ de l'éducation, développant des approches adéquates pour étudier des programmes dans ce domaine. Ses travaux avec Melvin M. Mark ont permis de clarifier ce que pouvait être une évaluation réaliste en pratique. Il s'est aussi attaché à comprendre quelle pouvait être l'influence de l'évaluation dans des sociétés démocratiques, en en exposant les mécanismes.

Ernest R. House est professeur émérite en éducation à l'Université du Colorado à Boulder. Il a commencé sa carrière en évaluant des programmes dans le champ de l'éducation. Il cherche très tôt à articuler son activité évaluative avec les questions de justice sociale, pour en faire un vecteur de diffusion des valeurs démocratiques, y compris contre toutes les formes de populisme. Il amène également des réflexions sur la pratique évaluative et la prise en compte des valeurs dans l'évaluation.

Marthe Hurteau, Ph. D. est professeure titulaire à l'Université du Québec à Montréal (UQAM) et responsable d'un diplôme supérieur en évaluation de programme à l'École des sciences de la gestion (UQAM). Elle a obtenu le titre d'évaluatrice qualifiée (EQ) décerné par la Société canadienne d'évaluation et elle est membre du jury d'accréditation. Parmi ses champs d'intérêt en recherche qui ont donné lieu à des publications scientifiques

: le jugement crédible en évaluation, les relations avec les parties prenantes, la contribution de l'intuition ainsi que de la sagesse pratique au sein de la pratique, et tout récemment, de la pratique éthique.

Steve Jacob est professeur de science politique à l'Université Laval au Québec. Il s'attache à analyser les processus de modernisation des administrations et la mise en place de processus de gestion de la performance. Ses travaux ont notamment porté sur l'institutionnalisation de l'évaluation dans plusieurs pays.

Farah Jamal (1986-2016) a été chercheuse à l'institut d'éducation de *University college London*, spécialisée en éducation et en santé des jeunes.

Karen E. Kirkhart est professeure émérite à l'Université de Syracuse, spécialisée en évaluation. Elle est particulièrement connue pour ses travaux dans lesquels elle reconceptualise la notion d'utilisation de l'évaluation en influence. Ses travaux ont également porté sur l'évaluation attentive aux différences culturelles. Elle introduit l'idée de validité multiculturelle, la validité de l'évaluation étant considérée comme enracinée dans un contexte culturel situé.

Yvonna S. Lincoln est professeure émérite en éducation et ressources humaines à l'Université Texas A&M. Elle met en avant, dans ses travaux, l'incapacité de la « méthode scientifique » classique à répondre de façon satisfaisante aux questions évaluatives et à réagir aux besoins et attentes de celles et ceux qui sont directement concerné-e-s par l'évaluation. C'est ce qui l'amène, avec Egon G. Guba, à proposer une approche constructiviste de l'évaluation, l'investigation naturaliste.

Ana Manzano est professeure associée en politiques publiques à l'Université de Leeds. Spécialiste en évaluation réaliste, elle s'intéresse notamment à l'évaluation des interventions complexes en santé.

Bruno Marchal est professeur associé à l'institut de médecine tropicale de l'Université d'Anvers, où il dirige l'unité « complexité et santé ».

Melvin M. Mark est professeur de psychologie à l'Université de l'État de Pennsylvanie. Il s'est particulièrement penché dans sa carrière sur les usages des sciences sociales, et notamment de la psychologie, dans l'évaluation. Il est aussi un des auteurs clés de l'évaluation réaliste, précisant la notion de mécanisme avec Gary T. Henry. Il se situe dans la suite de Karen Kirkhart pour parler d'influence de l'évaluation, dont il décortique justement les mécanismes.

Sandra Mathison est professeure en éducation à l'Université de Colombie Britannique (UBC). Elle a contribué à la théorie et la pratique des méthodes qualitatives en évaluation. Elle a notamment précisé ce qu'on appelait « triangulation » en évaluation. Elle travaille notamment sur le champ de l'éducation et a étudié en particulier l'usage des tests standardisés et leurs implications sociales, politiques et éthiques. Elle porte une voix critique sur l'évaluation dans une société néo-libérale.

Joseph A. Maxwell est professeur en éducation à l'Université George Mason. Ses apports se situent essentiellement dans le champ des méthodes qualitatives, notamment dans l'évaluation, et leur intégration avec des méthodes quantitatives.

Donna Mertens est professeure émérite en évaluation à l'Université Gallaudet. Elle a exploré pendant une longue partie de sa carrière le paradigme transformationnel et ses conséquences en termes de rôle de l'évaluateur/-trice et de pratique évaluative. Par rapport à Ernest House, elle insiste sur la nécessité de nommer les maux – racisme, sexisme, etc. – et de donner une place centrale à l'expérience vécue par celles et ceux qui subissent les discriminations.

Lawrence B. Mohr est professeur émérite de science politique et de politiques publiques à l'Université du Michigan. Il est spécialisé en évaluation, en méthodologie de la recherche et en épistémologie.

Nathalie Mons est sociologue, professeure du CNAM, titulaire de la chaire Évaluation des politiques éducatives, et responsable du Centre national d'étude des systèmes scolaires (Cnesco). Elle est spécialisée dans l'analyse des politiques éducatives française et étrangères.

Michael Morris est professeur émérite de psychologie à l'Université de New Haven. Il se spécialise sur les questions d'éthique en évaluation.

Mark Pearson est directeur adjoint chargé de l'emploi, du travail et des affaires sociales à l'OCDE.

Robert Picciotto a fait l'essentiel de sa carrière dans les institutions internationales. Il a notamment dirigé le groupe d'évaluation indépendant de la Banque mondiale, avant de s'orienter vers le conseil et la formation. Ses réflexions portent essentiellement sur l'évaluation du développement.

Michael Q. Patton a fait de l'utilité des travaux une pierre angulaire de l'évaluation et le critère ultime pour juger de leur qualité, et a proposé des approches nouvelles pour le faire. Auteur prolifique, il plaide notamment pour une redéfinition du rôle de l'évaluateur/-trice et des cadres classiques de l'évaluation pour accompagner les innovations sociales.

Anne Revillard est professeure associée en sociologie à Sciences Po, membre de l'Observatoire sociologique du changement (OSC) et directrice du Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP).

Valéry Ridde est directeur de recherche au CEPED (http://www.ceped.org), une unité de recherche commune à l'Université de Paris et à l'Institut de recherche pour le développement (IRD). Il a été professeur associé à l'École de santé publique de l'Université de Montréal. Il est actuellement basé à l'Institut de la santé et du développement (ISED) de l'Université Cheikh Anta Diop de Dakar (Sénégal).

Jim Rugh a travaillé toute sa vie dans le monde du développement international, en tant que consultant puis au sein d'une ONG.

Thomas A. Schwandt est professeur en éducation à l'Université de l'Illinois. Il est un théoricien de l'évaluation. Il s'est interrogé sa carrière durant sur la façon de concilier, en théorie et en pratique, les aspects moraux, politiques et techniques de l'évaluation. En amenant le concept de sagesse pratique appliqué à l'évaluation, il offre un nouveau cadre pour comprendre la pratique évaluative. Il a aussi contribué à clarifier les hypothèses sous-jacentes de méthodologies utilisées en évaluation.

Michael Scriven a eu des contributions très nombreuses dans le champ de l'évaluation, inventant un certain nombre des termes qui y sont employés (évaluation formative et sommative, par exemple). Il formalise la logique de l'évaluation, parfois considérée comme la seule réelle théorie de l'évaluation (Alkin), dans laquelle il affirme l'importance du jugement, à rebours des approches techniques alors dominantes. Il a également proposé des approches telles que l'évaluation affranchie des objectifs ou des méthodes telles que le modus operandi pour tester des relations causales.

Nichola Shackleton est chercheuse à l'Université d'Auckland, spécialisée dans l'étude des inégalités de santé chez les enfants.

Willliam Shadish (1949-2016), professeur de psychologie et de statistiques à l'Université de Californie à Merced, a contribué à constituer un corpus théorique sur l'évaluation, en collaboration notamment avec Thomas Cook. Ses travaux l'ont également amené à travailler avec ce dernier sur les méthodes expérimentales et quasi-expérimentales en évaluation.

Daniel L. Stufflebeam (1936-2017), professeur émérite à l'Université de Western Michigan, a longtemps été le président du *Joint Committee on Standards for Education Education* étatsunien et a fondé un centre de recherche portant sur l'évaluation en éducation. À ce titre il s'est largement penché sur les questions de critères et de tests standardisés. Il a développé un des premiers modèles d'évaluation de programme, le CIPP (Context, Input, Process and Product).

Edward A. Suchman soutient la nécessité d'utiliser la logique scientifique dans l'évaluation, tout en prenant en compte les contraintes pratiques à la démarche évaluative. Ses travaux portèrent en particulier sur le champ de la santé publique.

Sandy Taut est cheffe d'unité à l'agence qualité du service de l'éducation de l'État de Bavière. Psychologue, elle est experte des aspects théoriques et méthodologiques de l'assurance qualité dans l'éducation. Elle propose des conseils et de la formation à l'évaluation interne et externe et s'intéresse aux échanges entre la science et la pratique.

Sara van Belle est chercheuse en santé publique à l'Institut de médecine tropicale (ITM) d'Anvers. Politiste et anthropologue, elle est spécialisée en évaluation réaliste dans le domaine des politiques de développement.

Russell Viner est professeur en santé des adolescents à *University College London*, spécialisé dans l'évaluation des interventions en santé en direction des enfants et des jeunes.

Dana Wanzer est professeure assistante en psychologie à l'Université du Wisconsin, et pratique parallèlement en tant que consultante en évaluation. Elle évalue en particulier des programmes dans le champ de l'éducation.

Emily Warren est consultante en évaluation et doctorante en psychologie sociale à *Claremont Graduate University*.

Nan Wehipeihana est évaluatrice. Elle s'est particulièrement attachée à faire le lien dans son travail avec la culture maori et à ouvrir des espaces de dialogue qui facilitent la compréhension des Māori.

Carol H. Weiss (1927-2013), professeure en éducation à Harvard, a eu un rôle pionnier pour faire reconnaître la dimension politique de l'évaluation. Elle élargit la notion d'usage de l'évaluation et amène à reconsidérer l'évaluation comme un facteur parmi d'autres affectant la décision, qui se forme progressivement. Les résultats des évaluations percolent

progressivement et peuvent in fine changer la perspective des acteurs sur les problèmes à résoudre et leurs solutions. Elle a également un rôle important dans la construction d'une évaluation basée sur la théorie.

Gill Westhorp est chercheuse à l'Université Charles Darwin, où elle dirige la *Realist Research Evaluation and Learning Initiative* (RREALI). Elle est spécialisée en évaluation réaliste.

Howard White s'inscrit dans le courant de la politique fondée sur des données probantes. Après une carrière académique et à la Banque mondiale, il a dirigé l'Initiative internationale pour l'évaluation d'impact (3ie) puis la *Campbell Collaboration*. Il a notamment travaillé à réconcilier approches basées sur la théorie et évaluation contrefactuelle.

Geoff Wong est professeur associé en soins primaires à l'Université de Nuffield, spécialisé en évaluation réaliste.

Remerciements

Ce projet d'anthologie de textes fondamentaux en évaluation, sur une idée originale de Valéry Ridde, s'est concrétisé grâce au soutien financier et logistique du Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP). Il a ainsi bénéficié du soutien apporté par l'ANR et l'État au titre du programme d'investissements d'avenir dans le cadre du LABEX LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02) et de l'IdEx Université de Paris (ANR-18-IDEX-0001). Le projet a également bénéficié du soutien financier de la Chaire Innovation Publique.

Nous remercions vivement l'équipe du LIEPP pour l'appui apporté à ce projet à toutes ses étapes : Bruno Palier qui, alors directeur du LIEPP, a accueilli et soutenu avec enthousiasme cette initiative; Andreana Khristova pour tout le travail de coordination, de relations avec l'éditeur, et sa vigilance dans la mise en forme du manuscrit; Samira Jebli qui a effectué toutes les démarches de demande de renseignements et d'achat des droits de traduction auprès des éditeurs anglophones, et a contribué à la mise en forme du manuscrit; Sofia Cerdá Aparicio pour son appui essentiel à la préparation des fichiers et au formatage des textes.

Nous remercions aussi chaleureusement les associé-e-s de la Scop Quadrant Conseil qui ont soutenu ce projet, participé à la relecture de chapitres et accepté que deux de leurs collègues s'investissent pleinement dans ce projet.

Carine Gazier a fourni un travail considérable de traduction initiale des textes avant reprise par les coordinateurs et coordinatrices de l'ouvrage. Elle a également saisi toutes les références bibliographiques sous Zotero. Nous la remercions sincèrement pour son implication dans ce projet, sa disponibilité et la rigueur de son travail. Edgard Dewitte et Léo Le Roux ont fourni une aide essentielle de dernière minute au projet en travaillant respectivement sur la mise en forme des schémas et tableaux et sur

l'homogénéisation des modalités d'écriture inclusive. Merci à Sylvie Robel pour ses relectures minutieuses et son appui au lissage de l'expression, et à Alexandre Prince pour la saisie du manuscrit sous Pressbooks.

Enfin, nous adressons de vifs remerciements aux Éditions science et bien commun, et tout particulièrement Erika Nimis, pour avoir accepté ce projet de publication, ainsi que pour leur souplesse et leur réactivité dans sa prise en charge. Nous nous réjouissons de rejoindre avec cette anthologie ce beau projet collectif de diffusion scientifique en accès libre.

Premiers retours sur l'ouvrage

« L'évaluation sert à constater la valeur de biens publics. Ce livre introduit la francophonie au progrès déjà réalisé dans le monde académique anglo-saxon sur le chemin vers une pratique qui évalue mieux la conception des programmes sociaux, la qualité de leur mise en œuvre, et les effets qui leur sont attribués, sans oublier l'utilisation des résultats empiriques issus de l'évaluation. On cherche en vain un texte en français sur l'évaluation de la qualité de celui-ci. Lisez-le! Vous y découvrirez un nouveau corpus à la fois intellectuel et pratique, qui pourrait même améliorer les services publics offerts dans votre pays. »

Thomas D. Cook, professeur émérite de sociologie à l'Université de Northwestern et membre émérite de leur *Institute for Policy Research*

« Cette anthologie remplit un vide, et en même temps offre une perspective. Elle paraît au moment où l'on voudrait mieux connaître les effets de politiques conçues pour résoudre des problèmes toujours plus complexes, et quand forte est l'illusion de pouvoir se concentrer sur des pratiques, toutes sophistiquées qu'elles soient. Et pousse à chérir une longue expérience au cours de laquelle on a raisonné autour de problèmes cruciaux se référant aux usages, aux méthodes, aux tendances de l'évaluation.

Cette activité théorique-pratique est née et s'est diffusée dans un environnement (les États-Unis) et à un moment historique (le lancement des politiques de *welfare* des années 1960) différents de ceux d'aujourd'hui, mais elle a intéressé d'autres environnements, et d'autres champs d'intervention, grâce à sa versatilité et à sa capacité de comprendre les exigences de connaissance et de développement répandues partout. Par conséquent, la traduction de ces textes – "fondateurs" et contemporains – loin d'être une opération purement académique, offre l'opportunité d'un éclairage théorique sur des questions qui ont une forte

pertinence pratique dans le monde actuel, et qui peuvent fournir une précieuse orientation à tous les acteurs et actrices qui sont impliqués : des décideurs et décideuses aux opérateurs et opératrices, des évaluateurs et évaluatrices à celles et ceux auxquels les politiques sont adressées et qui les font vivre.

À l'intérieur de chaque partie de l'anthologie sont présentés les arguments les plus importants autours desquels se sont déroulés les débats qui ont fait progresser les connaissances évaluatives. Ici, les auteurs et autrices ont proposé un regard complet et non partisan, tout en opérant un choix inspiré des critères fortement cohérents avec les enjeux de l'évaluation : pluralisme (des éléments de définition et des approches), démocratie, aspiration à contribuer à un monde plus juste.Un mérite non moins important de cette anthologie est sa disponibilité en *open access*: un service à la démocratie. »

Nicoletta Stame, Université de Rome 'La Sapienza'

« Ce livre est une œuvre magistrale et intemporelle, qui illumine tant dans sa forme que dans le propos. Ouvrage de référence traitant tant de l'épistémologie que de la praxis de l'évaluation, il vient pallier le manque d'accès par le public francophone, aux classiques dans un domaine longtemps emporté par la marée de l'anglicisation. »

Sanni Yaya, professeur titulaire, vice-recteur International et Francophonie, Université d'Ottawa

À propos des Éditions science et bien commun

Les Éditions science et bien commun sont une branche de l'Association science et bien commun (ASBC), un organisme sans but lucratif enregistré au Québec depuis juillet 2011.

L'Association science et bien commun

L'Association science et bien commun se donne comme mission d'appuyer et de diffuser des travaux de recherche transuniversitaire favorisant l'essor d'une science pluriverselle, ouverte, juste, plurilingue, non sexiste, non raciste, socialement responsable, au service du bien commun.

Pour plus d'information, écrire à info@scienceetbiencommun.org, s'abonner à son compte Twitter @ScienceBienComm ou à sa page Facebook : https://www.facebook.com/scienceetbiencommun

Les Éditions science et bien commun

Un projet éditorial novateur dont les principales valeurs sont les suivantes.

- la publication numérique en libre accès, en plus des autres formats
- la pluridisciplinarité, dans la mesure du possible
- le plurilinguisme qui encourage à publier en plusieurs langues, notamment dans des langues nationales africaines ou en créole, en plus du français

- l'internationalisation, qui conduit à vouloir rassembler des auteurs et autrices de différents pays ou à écrire en ayant à l'esprit un public issu de différents pays, de différentes cultures
- mais surtout la justice cognitive :
    - chaque livre collectif, même s'il s'agit des actes d'un colloque, devrait aspirer à la parité entre femmes et hommes, entre juniors et seniors, entre auteurs et autrices issues du Nord et issues du Sud (des Suds); en tout cas, tous les livres devront éviter un déséquilibre flagrant entre ces points de vue;
    - chaque livre, même rédigé par une seule personne, devrait s'efforcer d'inclure des références à la fois aux pays du Nord et aux pays des Suds, dans ses thèmes ou dans sa bibliographie;
    - chaque livre devrait viser l'accessibilité et la « lisibilité », réduisant au maximum le jargon, même s'il est à vocation scientifique et évalué par les pairs.

Le catalogue

Le catalogue des Éditions science et bien commun (ESBC) est composé de livres qui respectent les valeurs et principes des ÉSBC énoncés ci-dessus.

- Des ouvrages scientifiques (livres collectifs de toutes sortes ou monographies) qui peuvent être des manuscrits inédits originaux, issus de thèses, de mémoires, de colloques, de séminaires ou de projets de recherche, des rééditions numériques ou des manuels universitaires. Les manuscrits inédits seront évalués par les pairs de manière ouverte, sauf si les auteurs ne le souhaitent pas (voir le point de l'évaluation ci-dessus).
- Des ouvrages de science citoyenne ou participative, de vulgarisation scientifique ou qui présentent des savoirs locaux et patrimoniaux, dont le but est de rendre des savoirs accessibles au plus grand nombre.
- Des essais portant sur les sciences et les politiques scientifiques (en études sociales des sciences ou en éthique des sciences, par

exemple).

- Des anthologies de textes déjà publiés, mais non accessibles sur le web, dans une langue autre que le français ou qui ne sont pas en libre accès, mais d'un intérêt scientifique, intellectuel ou patrimonial démontré.
- Des manuels scolaires ou des livres éducatifs pour enfants

Pour l'accès libre et universel, par le biais du numérique, à des livres scientifiques publiés par des autrices et auteurs de pays des Suds et du Nord

Pour plus d'information : écrire à info@editionscienceetbiencommun.org