

MÉTHODES ET APPROCHES EN
ÉVALUATION
DES POLITIQUES PUBLIQUES

SOUS LA DIRECTION DE
ANNE REVILLARD

Méthodes et approches en évaluation des politiques publiques

Sous la direction d'Anne Revillard

QUÉBEC : ÉDITIONS SCIENCE ET BIEN COMMUN



Méthodes et approches en évaluation des politiques publiques Copyright © by Anne Revillard
is licensed under a License Creative Commons Attribution - Partage dans les mêmes
conditions 4.0 International, except where otherwise noted.

Titre : Méthodes et approches en évaluation des politiques publiques

Sous la direction d'Anne Revillard

Design de la couverture : Kate McDonnell

Édition et révision linguistique : Alizée Harel et Érika Nimis

ISBN pour l'impression : 978-2-925128-27-4

ISBN pour le PDF : 978-2-925128-28-1

Dépôt légal – Bibliothèque et Archives nationales du Québec 2023

Dépôt légal – Bibliothèque et Archive nationale Canada

Ce livre est publié sous licence Creative Commons CC BY-SA 4.0 et disponible en
libre accès à <https://scienceetbiencommun.pressbooks.pub/evaluationpolpub/>

Éditions science et bien commun

<http://editionscienceetbiencommun.org>

3-855 avenue Moncton

Québec (Québec) G1S 2Y4

Diffusion : info@editionscienceetbiencommun.org

Table des matières

In memoriam	ix
Introduction	1
Glossaire	25
Partie I. MÉTHODES QUANTITATIVES	
1. Essais contrôlés randomisés	37
<i>Carlo Barone</i>	
2. Méthode des doubles différences (difference-in-differences)	47
<i>Denis Fougère et Nicolas Jacquemet</i>	
3. La régression sur discontinuité	61
<i>Denis Fougère et Nicolas Jacquemet</i>	
4. Méthodes d'appariement	73
<i>Pauline Givord</i>	
5. Microsimulation	83
<i>Mathias André</i>	
6. Expérimentation en laboratoire	95
<i>Lou Safra</i>	
7. Testing	107
<i>Nicolas Jacquemet</i>	
8. L'analyse coût-efficacité	121
<i>Thomas Rapp</i>	

Partie II. MÉTHODES QUALITATIVES

9. Observation directe et ethnographie	133
<i>Nicolas Fischer</i>	
10. L'entretien semi-directif	145
<i>Clément Pin</i>	
11. Les focus groups	157
<i>Ana Manzano</i>	
12. Entretiens de groupe	169
<i>Charlotte Halpern</i>	
13. Les études de cas	181
<i>Valéry Ridde, Abdourahmane Coulibaly, Lara Gautier</i>	
14. Traçage de processus	193
<i>Estelle Raimondo</i>	
15. L'analyse historique comparée	205
<i>Emanuele Ferragina</i>	

Partie III. MÉTHODES MIXTES ET APPROCHES TRANSVERSALES

16. Les méthodes mixtes	221
<i>Pierre Pluye</i>	
17. Les revues de littérature systématiques mixtes	233
<i>Quan Nha Hong</i>	
18. Les comparaisons de niveau macro	245
<i>Emanuele Ferragina</i>	
19. L'analyse qualitative comparée	257
<i>Valérie Pattyn</i>	
20. L'évaluation basée sur la théorie	269
<i>Agathe Devaux-Spatarakis</i>	

21. Évaluation réaliste	279
<i>Sarah Louart, Habibata Baldé, Émilie Robert et Valéry Ridde</i>	
22. L'analyse de contribution	293
<i>Thomas Delahais</i>	
23. Récolte d'incidences	307
<i>Genowefa Blundo Canto</i>	
24. Sécurisation culturelle	319
<i>Loubna Belaid et Neil Andersson</i>	
Présentation des auteur·rice·s	337
Remerciements	351
À propos des Éditions science et bien commun	353

Cet ouvrage est dédié à la mémoire du Professeur Pierre Pluye,
qui a tant apporté à la réflexion sur les méthodes mixtes.

Introduction

Cette publication fait suite à un premier ouvrage (*Évaluation : fondements, controverses, perspectives*) publié fin 2021 aux éditions Science et bien commun (ESBC) avec l'appui du Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP), compilant une série d'extraits de textes fondamentaux et contemporains en évaluation (Delahais et al. 2021). Bien qu'une partie de ce livre soit dédiée à la diversité des approches paradigmatiques, nous avons choisi de ne pas rentrer dans le détail de la présentation des méthodes au motif que celle-ci mériterait au moins un ouvrage à part entière. C'est l'objet de ce volume. Cette publication s'inscrit doublement dans le projet collectif du LIEPP, par l'articulation proposée entre recherche et évaluation, et par l'ouverture d'un dialogue entre méthodes quantitatives et qualitatives.

Des méthodes entre recherche et évaluation

La plupart des définitions de l'évaluation de programmes¹ articulent trois dimensions, décrites par Alkin et Christie comme les trois branches de « l'arbre des théories en évaluation » (Alkin et Christie 2012) : la mobilisation de *méthodes* de recherche (l'évaluation prend appui sur une démarche d'investigation empirique systématique), le rôle joué par les

1. Citons par exemple la définition proposée par Michaël Patton, caractérisant l'évaluation comme « la collecte systématique d'informations sur les activités, caractéristiques et effets des programmes, dans le but de développer des jugements sur ceux-ci, d'améliorer leur efficacité, et d'informer les décisions futures relatives à ce programme » (*“the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgements about the program, improve program effectiveness, and/or inform decisions about future programming”*) (Patton 1997, 23).

valeurs, au fondement des critères à partir desquels on va porter un jugement sur l'intervention étudiée, et l'attention prêtée à l'utilité de l'évaluation.

Le fait de recourir à des méthodes systématiques d'investigation empirique constitue donc un des fondements de la pratique de l'évaluation. C'est ce en quoi l'évaluation au sens de recherche évaluative se distingue du simple jugement subjectif que peut par ailleurs désigner le terme « évaluation » dans son sens commun (Suchman 2021). L'évaluation est d'abord une pratique de recherche appliquée et, à ce titre, elle a emprunté toute une série de techniques d'enquête, aussi bien quantitatives que qualitatives, initialement développées en recherche fondamentale (ex. questionnaires, analyses quantitatives sur des bases de données, méthodes expérimentales, entretiens semi-directifs, observations, étude de cas, etc.). Au-delà des techniques, l'emprunt concerne aussi les modalités d'analyse et la conception des designs de recherche. En dépit de ce fort lien méthodologique, l'évaluation ne se réduit pas à une pratique de recherche (Wanzer 2021), comme le suggèrent les deux autres dimensions précédemment identifiées (souci des valeurs et de l'utilité). De fait, le développement de l'évaluation de programmes a donné lieu à une pluralité de pratiques par une diversité d'acteurs publics et privés (administrations publiques, consultant-e-s, O.N.G, etc.), pratiques au sein desquelles les enjeux méthodologiques ne sont pas nécessairement centraux et où la rigueur méthodologique peut être très variable.

Parallèlement, la pratique de l'évaluation est restée faiblement et très inégalement institutionnalisée à l'université (Cox 2021), où elle souffre notamment d'une fréquente dévalorisation des pratiques de recherche appliquée, de soupçons de complaisance vis-à-vis des commanditaires, ou encore de difficultés liées à son caractère interdisciplinaire (cf *infra*) (Jacob 2021). Ainsi, bien qu'elle ait développé ses revues et ses conférences professionnelles, l'évaluation fait encore l'objet de peu de programmes doctoraux et de recrutements dédiés. Pratiquée à des degrés divers par différentes disciplines universitaires (la santé publique, l'économie et le

développement sont aujourd'hui particulièrement impliqués), et parfois qualifiée de « transdiscipline » sur le plan de sa portée épistémologique (Scriven 2021), l'évaluation est encore loin de constituer une discipline universitaire au sens institutionnel du terme. Sur le plan épistémologique, on peut se féliciter de cette non- (ou faible) disciplinarisation de l'évaluation à l'Université. Il n'en demeure pas moins que celle-ci induit des fragilités. Une des conséquences de cette situation est un fréquent défaut de formation des chercheurs et chercheuses en évaluation : notamment concernant les dimensions non méthodologiques de cette pratique (questions des valeurs et de l'utilité), mais aussi quant à certaines méthodes plus spécifiquement issues de l'évaluation.

En effet, si l'évaluation a largement emprunté aux méthodes de sciences sociales, elle a réciproquement été le creuset d'un certain nombre d'innovations méthodologiques. C'est ainsi à partir de la démarche d'évaluation, initialement en éducation dans les années 1920 puis dans les politiques sociales, la santé et d'autres domaines à partir des années 1960, que l'usage de méthodes expérimentales a pris son essor dans les sciences sociales (Campbell et Stanley 1963). Le rapprochement avec la médecine occasionné par l'emprunt du modèle de l'essai clinique (les notions « d'essai » et de « traitement » ayant ainsi été transposées à l'évaluation) a ensuite favorisé le transfert des sciences médicales à l'évaluation d'une autre méthode, les revues systématiques de littérature, consistant à se doter d'un protocole systématique pour rechercher les publications existantes sur une (ou des) question(s) évaluative(s) donnée(s) et dresser une synthèse de leurs apports (Hong et Pluye 2018; Belaid et Ridde 2020). Sans en être le seul lieu de déploiement, l'évaluation de programmes a aussi fortement contribué au développement et à la théorisation des méthodes mixtes, consistant à articuler des techniques d'enquêtes et/ou démarches d'analyse qualitatives et quantitatives dans une même recherche (Baiz et Revillard 2022; Greene, Benjamin, et Goodyear 2001; Burch et Heinrich 2016; Mertens 2017). De façon similaire, du fait de sa préoccupation centrale vis-à-vis de l'utilisation des savoirs,

l'évaluation a été un lieu privilégié du développement des recherches participatives et de leur théorisation (Brisolara 1998; Cousins et Whitmore 1998; Patton 2018).

Si ces méthodes (méthodes expérimentales, revues systématiques de littérature, méthodes mixtes, recherche participative) sont immédiatement applicables à d'autres domaines que l'évaluation de programmes, d'autres approches et outils méthodologiques ont été plus spécifiquement développés à cette fin². C'est notamment le cas des évaluations basées sur la théorie (Weiss 1997; Rogers et Weiss 2007), incluant une diversité d'approches (évaluation réaliste, analyse de contribution, récolte d'incidence...) qui seront décrites ci-dessous (Pawson et Tilley 1997; Mayne 2012; Wilson-Grau 2018). En dehors de quelques disciplines au sein desquelles elles sont plus diffusées, comme la santé publique ou le développement (Ridde et Dagenais 2009; Ridde et al. 2020), ces approches restent encore peu connues des chercheurs et chercheuses ayant suivi des formations classiques en méthodes de recherche... Y compris de celles et ceux pouvant être amené·e·s à s'impliquer dans des projets d'évaluation.

Un dialogue a donc besoin d'être renoué entre évaluation et recherche : selon la dynamique réciproque de l'emprunt initial aux méthodes de recherche par l'évaluation, une plus grande diversité de milieux en recherche fondamentale gagne aujourd'hui à mieux connaître les méthodes et approches spécifiques développées à partir de la pratique de l'évaluation. C'est une des vocations du LIEPP, qui promeut un renforcement des échanges entre chercheurs et praticien·ne·s de l'évaluation. Le LIEPP organise ainsi depuis 2020 un séminaire mensuel sur les méthodes et approches en évaluation (METHEVAL) alternant des

2. Par distinction avec les méthodes au sens d'outils méthodologiques, les approches se situent dans « une sorte d'entre-deux entre la théorie et la pratique » (Delahais 2022), en incarnant certains paradigmes. En évaluation, certaines peuvent être très orientées sur la dimension méthodologique, mais d'autres plus concernées par les valeurs, par l'utilisation des résultats, ou encore par la justice sociale (ibid.).

présentations de chercheurs et de praticien-ne-s, et réunissant une diversité de public³. C'est également une des motivations à l'origine de l'ouvrage *Évaluation : fondements, controverses, perspectives*, publié en 2021, qui visait notamment à sensibiliser les chercheurs-euses aux aspects non méthodologiques de l'évaluation (Delahais et al. 2021). La présente publication vient compléter la démarche en facilitant l'appropriation d'approches développées en évaluation telle que l'évaluation basée sur la théorie, l'évaluation réaliste, l'analyse de contribution et la récolte d'incidence (*outcome harvesting*).

Mais réciproquement, le LIEPP fait le pari que l'évaluation gagne à davantage s'ouvrir à des outils méthodologiques plus fréquemment mobilisés dans la recherche fondamentale et dont elle est moins familière, notamment du fait du ciblage du questionnement à l'échelle de l'intervention. En effet, l'évaluation prend classiquement pour objet une intervention ou un programme, à une échelle le plus souvent locale, régionale ou nationale, et dans un périmètre de questionnement suffisamment ciblé pour permettre de tirer des conclusions quant aux conséquences de l'intervention étudiée. Si nous parlons d'évaluation des politiques publiques plutôt que d'évaluation de programmes au sens strict, c'est pour inclure la possibilité d'une réflexion à une échelle plus macro au sens à la fois géographique et temporel, en intégrant des réflexions sur l'historicité des politiques publiques, sur l'agencement de différentes interventions dans un contexte plus large d'action publique (tel ou tel régime d'État-providence par exemple), et en s'appuyant plus systématiquement sur des démarches de comparaison internationale. Il s'agit donc de relier l'évaluation à une démarche d'analyse des politiques publiques, comme le défendaient déjà dans les années 1990 les promoteurs d'une « évaluation à la française » (Duran, Monnier, et Smith 1995; Duran, Erhel, et Gautié 2018). C'est ce que permettent par exemple

3. Le programme ainsi que des ressources issues des séances précédentes de ce séminaire sont disponibles en ligne : <https://www.sciencespo.fr/liepp/fr/content/cycle-de-seminaires-methodes-et-approches-en-evaluation-metheval.html>

l'analyse historique comparée et les comparaisons de niveau macro, présentées dans cet ouvrage. Un autre implicite important de l'évaluation de programme est la focalisation du questionnement sur l'intervention étudiée. En décentrant le regard, de nombreuses pratiques de recherche fondamentale peuvent apporter des compléments de réflexion très utiles de façon plus prospective, en aidant à la compréhension des problèmes sociaux ciblés par les interventions. Toutes les recherches thématiques menées en sciences sociales apportent à cet égard des éclairages très utiles aux pratiques d'évaluation (Rossi, Lipsey, et Freeman 2004). Parmi les méthodes présentées dans cet ouvrage, des approches expérimentales comme l'expérimentation en laboratoire ou les démarches de testing, non nécessairement centrées sur les interventions en tant que telles, permettent d'illustrer cet apport plus prospectif de la recherche à l'évaluation.

Un dialogue entre approches qualitatives et quantitatives

En empruntant aux sciences sociales ses méthodes, l'évaluation des politiques publiques a hérité des controverses méthodologiques et épistémologiques associées. Bien que les appels à la réconciliation soient nombreux, que l'évaluation mette plus volontiers en avant son pragmatisme méthodologique (la question évaluative guide le choix des méthodes), et qu'elle ait joué un rôle moteur dans le développement des méthodes mixtes, dans la pratique, en évaluation comme en recherche, le dialogue entre traditions quantitative et qualitative (surtout dans leur dimension épistémologique) n'est pas toujours simple.

Articuler différentes approches disciplinaires et méthodologiques pour évaluer les politiques publiques est l'ambition fondatrice du LIEPP. Les difficultés de ce dialogue, notamment sur le plan épistémologique (opposition entre positivisme et constructivisme) ont été identifiées dès

la création du laboratoire (Wasmer et Musselin 2013). Au fil des années, le LIEPP s'est efforcé de surmonter ces obstacles en organisant le dialogue entre différentes méthodes et disciplines pour enrichir l'évaluation : par le développement de six axes de recherche coanimés par des chercheurs et chercheuses de différentes disciplines, par la conduite de projets menés par des équipes interdisciplinaires, mais aussi par la mise en discussion régulière de projets issu d'une discipline ou famille de méthodes par des spécialistes d'autres disciplines ou méthodes. C'est aussi au fil de ces échanges qu'a émergé le besoin de supports didactiques permettant de faciliter la compréhension des méthodes quantitatives par les spécialistes de méthodes qualitatives, et vice-versa, compréhension qui devient de plus en plus difficile dans un contexte de technicisation croissante des méthodes. Cet ouvrage vient répondre à ce besoin, en prenant fortement appui sur le collectif de chercheuses et de chercheurs ouvert·e·s à l'interdisciplinarité et au dialogue entre méthodes qui s'est ainsi constitué au LIEPP au fil des années : sur les 25 autrices et auteurs de cet ouvrage, neuf sont affilié·e·s au LIEPP et huit autres ont eu l'occasion de présenter leurs travaux lors de séminaires organisés par le LIEPP.

Cet ouvrage a donc été conçu comme un moyen de favoriser le dialogue entre méthodes, au LIEPP et au-delà. L'objectif n'est pas nécessairement de promouvoir le développement de recherches par méthodes mixtes, même nous décrivons les atouts de ces démarches (Partie III). Il s'agit d'abord de favoriser une compréhension mutuelle entre les différentes approches méthodologiques, de faire en sorte que les praticiennes et praticiens de méthodes qualitatives comprennent en quoi consiste l'apport complémentaire des méthodes quantitatives, leur portée et leurs limites, et réciproquement. Ce faisant, la démarche vise aussi à favoriser une plus grande réflexivité dans chaque pratique méthodologique, grâce à une meilleure prise de conscience de ce pour quoi une méthode est la plus adaptée et des questions pour lesquelles d'autres méthodes sont plus pertinentes. Tout en évitant une technicité excessive, il s'agit de rentrer dans le fonctionnement de chaque méthode pour comprendre concrètement ce qu'elle permet et ce qu'elle ne permet pas de faire. Nous

faisons le pari que cette approche pratique aidera à dépasser certains obstacles au dialogue entre méthodes liés à de grandes oppositions épistémologiques (positivisme *versus* constructivisme par exemple) qui ne se retrouvent pas nécessairement dans la pratique quotidienne de la recherche. Pour les étudiant·e·s et pour un public non universitaire (notamment chez des responsables publics ou associatifs qui peuvent avoir recours à des évaluations de programmes), il s'agit par ailleurs de favoriser une compréhension plus globale des apports et limites des différentes méthodes.

Loin de prétendre à l'exhaustivité, l'ouvrage se propose de présenter quelques exemples de trois grandes familles de méthodes ou approches : les méthodes quantitatives, les méthodes qualitatives, et les méthodes mixtes et approches transversales en évaluation⁴. Dans ce qui suit, nous présentons l'organisation générale de l'ouvrage et les différents chapitres en les intégrant dans une réflexion plus globale sur la distinction entre approches quantitatives et qualitatives. À un niveau très général, méthodes quantitatives et qualitatives se distinguent par la densité et l'étendue du type d'information qu'elles produisent : alors que les méthodes quantitatives permettent de produire une information limitée sur un grand nombre de cas, les méthodes qualitatives fournissent une information plus dense et contextualisée sur un nombre restreint de cas. Mais au-delà de ces caractéristiques descriptives, les deux familles de méthodes se distinguent aussi tendanciellement dans leur conception de la causalité. Il s'agit là d'une question centrale pour l'évaluation des

4. Ce faisant, il vient compléter d'autres ressources méthodologiques disponibles sous forme d'ouvrages (Ridde et Dagenais 2009; Ridde et al. 2020; Newcomer, Hatry, et Wholey 2015; Mathison 2005; Weiss 1998; Patton 2015), ou encore en ligne : citons par exemple le *Methods excellence network* (<https://www.methodsnet.org/>), ou encore en évaluation, les ressources compilées par l'OCDE (<https://www.oecd.org/fr/cad/evaluation/keydocuments.htm>), le réseau Evalpartners de l'ONU (<https://evalpartners.org/>), ou encore en France les guides méthodologiques de l'Institut des politiques publiques (<https://www.ipp.eu/publications/guides-methodologiques-ipp/>) et de la Société coopérative et participative (SCOP) Quadrant Conseil (<https://www.quadrant-conseil.fr/ressources/evaluation-impact.php#/>).

politiques publiques qui, sans se restreindre à cette interrogation⁵, s'est fondée autour d'un questionnement sur l'impact des interventions publiques : dans quelle mesure peut-on imputer tel changement observé à l'effet de telle intervention ? – Soit une question causale (peut-on établir une relation de cause à effet entre l'intervention et le changement observé ?). Pour comprendre les apports complémentaires des méthodes quantitatives et qualitatives pour l'évaluation, il importe donc de saisir leurs différentes façons d'aborder cette question centrale de la causalité.

Méthodes quantitatives

Les méthodes quantitatives expérimentales et quasi-expérimentales reposent sur une conception contrefactuelle de la causalité. Selon cette conception, pour prouver que A est cause de B, il faut montrer que toutes choses égales par ailleurs, si A est absent, B est absent (Woodward 2003). Appliquée à l'évaluation de l'impact des politiques publiques, cette logique invite à prouver qu'une intervention provoque un impact donné en montrant qu'en l'absence de cette intervention, toutes choses égales par ailleurs, cet impact ne se produit pas (Desplat et Ferracci 2017). Toute la difficulté consiste alors à approximer du mieux possible ces situations « toutes choses égales par ailleurs » : que se serait-il passé en l'absence d'intervention, toutes les autres caractéristiques de la situation étant identiques ? C'est ce souhait de comparer des situations avec et sans intervention « toutes choses égales par ailleurs » qui a donné lieu au développement des méthodes expérimentales en évaluation (Campbell et Stanley 1963; Rossi, Lipsey, et Freeman 2004).

5. L'évaluation s'interroge aussi, par exemple, sur la pertinence, la cohérence, l'efficacité, l'efficience ou encore la soutenabilité des interventions. Voir OECD DAC Network on Development Evaluation (EvalNet) <https://www.oecd.org/dac/evaluation/dacriteriaforevaluatingdevelopmentassistance.htm>

La plupart des expérimentations menées en évaluation des politiques publiques sont des expérimentations de terrain (*field experiments*), au sens où elles étudient l'intervention en situation, telles qu'elle se déploie effectivement. Les essais (ou expérimentations) contrôlés randomisés (ECR, souvent désignés par l'acronyme RCT correspondant à la désignation anglophone de *randomised controlled trials*) (Cf. Chapitre 1) comparent ainsi un groupe expérimental (recevant l'intervention) et un groupe témoin en visant une équivalence de caractéristiques entre les deux groupes grâce à l'assignation aléatoire des participant·e·s à l'un ou l'autre groupe. Ce type d'approche est particulièrement adapté aux interventions par ailleurs désignées sous le terme « d'expérimentations » dans l'action publique (Devaux-Spatarakis 2014), c'est-à-dire des interventions que les pouvoirs publics lancent dans un nombre restreint de territoires ou d'organisations (ce qui permet l'existence de groupes témoins) pour en tester les effets⁶. Lorsque ce type d'expérimentation directe n'est pas possible, les évaluatrices et évaluateurs peuvent avoir recours à plusieurs méthodes quasi-expérimentales, visant à reconstituer des groupes de comparaisons à partir de situations et de données déjà existantes (donc sans manipulation du réel de leur part, contrairement aux protocoles expérimentaux) (Fougère et Jacquemet 2019). La méthode des doubles différences (couramment appelée « *diff-in-diff* » pour *difference-in-differences*) utilise un marqueur temporel à partir duquel un des deux groupes étudiés reçoit l'intervention et l'autre non, et mesure l'impact de l'intervention en comparant les résultats avant et après ce moment (Cf Chapitre 2). La régression sur discontinuité (Cf. Chapitre 3), quant à elle, reconstitue un groupe cible et un groupe de contrôle en comparant les situations de part et d'autre d'un seuil d'éligibilité fixé par la politique étudiée (par exemple, éligibilité à l'intervention à partir de

6. Ces initiatives des pouvoirs publics sont malheureusement loin de faire systématiquement l'objet d'évaluations rigoureuses.

tel âge, de tel seuil de revenu, etc.). Enfin, l'appariement (Cf. Chapitre 4) consiste à comparer les situations de bénéficiaires d'une intervention à celle de non bénéficiaires aux caractéristiques les plus proches possibles.

Parallèlement à ces méthodes qui prennent appui sur des données en situation réelle, d'autres démarches d'étude d'impact quantitatives prennent appui sur des simulations informatiques ou des expériences en laboratoire. La microsimulation (Cf. Chapitre 5), dont le développement a été favorisé par l'amélioration de la puissance de calcul informatique, consiste à estimer *ex ante* l'impact attendu d'une intervention en prenant en considération une grande variété de données relatives aux individus cibles et en simulant des changements de situation (par exemple, vieillissement, évolution du marché de l'emploi, politiques fiscales, etc.). Elle permet également d'étudier finement *ex post* la diversité des effets de politiques publiques sur les individus concernés. Mais l'évaluation des politiques publiques peut aussi prendre appui sur des expérimentations en laboratoire (Cf. Chapitre 6), permettant de mesurer de façon précise le comportement des individus, et notamment de mettre au jour des biais non conscients. De telles analyses peuvent par exemple être précieuses pour aider à la conception des politiques antidiscriminatoires, dans une démarche d'évaluation *ex ante*. C'est aussi dans le contexte de la réflexion sur ces politiques qu'a été développée la méthode d'études par correspondance, ou *testing* (Cf. Chapitre 7), permettant de mesurer les discriminations par l'envoi de candidatures fictives en réponse à des offres réelles (par exemple des offres d'emploi). Mais l'évaluation cherche aussi à mesurer l'efficacité des interventions, au-delà de leur impact. Ceci suppose de comparer les résultats obtenus au coût de la politique étudiée et à ceux de politiques alternatives, dans une démarche d'analyse coût – efficacité (Cf. Chapitre 8).

Méthodes qualitatives

Si elles sont par ailleurs compatibles avec des approches contrefactuelles, les méthodes qualitatives, quant à elles, viennent alimenter de façon privilégiée une conception plus générative ou processuelle de la causalité (Lawrence B. Mohr 2021; Maxwell 2004; 2012; Mohr 1999). Selon cette logique, la causalité est inférée, non plus de relations entre des variables, mais de l'analyse des processus par lesquels la causalité opère. Alors que l'approche contrefactuelle permet d'établir si A cause B, l'approche processuelle montre comment (par quelle série de mécanismes) A cause B, par l'observation des manifestations empiriques de ces mécanismes causaux qui relient A et B. Ce faisant, elle dépasse la logique behavioriste qui, dans les approches contrefactuelles, conçoit l'intervention selon une logique de stimulus-réponse, l'intervention en elle-même constituant alors une forme de boîte noire. À l'inverse, les approches plus qualitatives permettent de décomposer l'intervention en une série de processus qui contribuent à produire (ou à empêcher) le résultat escompté : c'est le principe général des évaluations basées sur la théorie dont la démarche, également compatible avec des méthodes quantitatives, est présentée dans la troisième partie de l'ouvrage (Cf. Chapitre 20). Cette analyse à une échelle plus fine est rendue possible par la focalisation sur un nombre limité de cas, qui sont alors étudiés de façon plus approfondie à partir de différentes techniques qualitatives. Une attention particulière est prêté aux contextes, ainsi qu'aux processus mentaux et aux logiques d'action des personnes impliquées dans l'intervention (agent·e·s chargé·e·s de sa mise en œuvre, publics cibles), dans une démarche compréhensive (Revillard 2018). À la différence des méthodes quantitatives, les méthodes qualitatives ne permettent pas de mesurer l'impact d'une politique publique; elles peuvent en revanche l'expliquer (et expliquer ses variations selon les contextes), mais aussi répondre à d'autres questions évaluatives telles que celle de la pertinence des interventions. Le tableau 1 résume ces différences tendanciennes entre méthodes quantitatives et qualitatives : il importe de préciser que nous mettons ici en évidence des affinités

privilégées de telle famille de méthodes avec telle approche de la causalité et telle prise en compte des processus et du contexte, mais il s'agit d'une distinction idéaltypique qui est loin d'épuiser les combinaisons possibles quant aux méthodes et conceptions de la recherche.

Méthodes	Approche de la causalité	Contexte	Mise en œuvre de la politique et processus mentaux des individus	Mesure de l'impact	Explication de l'impact
Quantitatives (expérimentales ou quasi-expérimentales)	Contrefactuelle : Pour prouver que A est cause de B, on montre que toutes choses égales par ailleurs, si A est absent, B est absent.	Efforts pour neutraliser son effet	Inconnus/ne peuvent faire l'objet que d'hypothèses (boîte noire)	Oui	Non
Qualitatives	Processuelle ou générative : Pour prouver comment A cause B, on observe et analyse la série de mécanismes qui relient A à B.	Intégré dans l'explication	Étudiés : une approche compréhensive, qui explore la subjectivité	Non	Oui

Tableau 1 : Méthodes quantitatives et qualitatives, deux approches différentes de la causalité

La technique d'enquête qualitative la plus emblématique est probablement l'observation directe ou ethnographie, héritière de l'anthropologie, qui consiste à observer directement sur le terrain la situation sociale que l'on cherche à étudier (Cf. Chapitre 9). Méthode particulièrement engageante, l'observation directe est très efficace pour mettre au jour tous ces processus intermédiaires de l'action publique qui contribuent à en produire les effets, ainsi que pour mettre à distance les discours officiels par l'observation directe des interactions. L'entretien semi-directif (Cf. Chapitre 10) est une autre technique d'enquête qualitative répandue, qui consiste en une interaction verbale sollicitée par le chercheur ou la chercheuse auprès d'un·e participant·e à la recherche, à partir d'une grille de questions utilisées de façon très souple. L'entretien vise à la fois la collecte d'information et la compréhension de l'expérience et de la vision du monde de la personne interviewée. Cette méthode peut également être déployée dans un cadre plus collectif, sous forme de

focus groups (Cf. Chapitre 11) ou d'entretiens de groupe (Cf. Chapitre 12). Comme le souligne Ana Manzano dans son chapitre sur les *focus groups*, les terminologies pour désigner ces pratiques d'entretiens collectifs sont variables. Notre objectif en publiant deux chapitres sur ces questions n'est pas de rigidifier la distinction mais d'apporter deux regards complémentaires sur ces méthodes qui sont très fréquemment utilisées en évaluation.

Bien que pouvant mobiliser une diversité de méthodes qualitatives, quantitatives et mixtes, les études de cas (Cf. Chapitre 13) s'inscrivent classiquement dans une tradition de recherche qualitative du fait de leur filiation avec l'anthropologie. Elles permettent d'étudier les interventions dans leur contexte, et elles sont particulièrement adaptés à l'analyse des interventions complexes. Plusieurs études de cas peuvent être combinées dans l'évaluation d'une même politique; les modalités de leur sélection sont alors décisives. Prenant principalement mais non exclusivement appui sur des techniques d'enquête qualitatives, le traçage de processus (*process tracing*) (Cf. Chapitre 14) se concentre sur le déroulement de l'intervention dans un cas particulier, en cherchant à retracer comment certaines actions en ont entraîné d'autres. L'évaluatrice ou l'évaluateur se comporte alors comme un-e détective recherchant les « empreintes digitales » laissées par les mécanismes de changement. La démarche permet d'établir dans quelles conditions, comment et pourquoi une intervention fonctionne dans un cas particulier. Enfin, l'analyse historique comparée combine les deux outils méthodologiques fondamentaux des sciences sociales que sont la comparaison et l'histoire pour aider à expliquer les phénomènes sociaux à grande échelle (Cf. Chapitre 15). Elle est particulièrement utile pour rendre compte de la définition des politiques publiques.

Méthodes mixtes et approches transversales en évaluation

La troisième et dernière partie de l'ouvrage regroupe une série de chapitres portant sur l'articulation entre méthodes qualitatives et quantitatives ainsi que sur des approches transversales en évaluation (compatibles avec une diversité de méthodes). L'évaluation des politiques publiques a joué un rôle moteur dans la formalisation du recours aux méthodes mixtes, conduisant notamment à la distinction entre différentes stratégies d'articulation entre méthodes qualitatives et quantitatives (devis séquentiel exploratoire, séquentiel explicatif ou convergent) (Cf. Chapitre 16). Même lorsque l'enquête empirique ne mobilise qu'un type de méthode, elle gagne à prendre appui sur une revue de littérature systématique mixte. Alors que la pratique des revues systématiques de littérature s'est initialement développée pour synthétiser les apports des travaux menés par essais contrôlés randomisés, cette pratique s'est diversifiée au fil des années pour inclure d'autres types de recherches (Hong et Pluye 2018). Les revues de littérature systématiques mixtes ont ainsi pour particularité d'intégrer des travaux à la fois quantitatifs, qualitatifs et mixtes, permettant de répondre à une plus grande diversité de questions évaluatives (Cf. Chapitre 17).

Une fois posé ce cadre général sur les méthodes et les revues mixtes, les chapitres suivants présentent six approches transversales. Les deux premières, les comparaisons de niveau macro et l'analyse qualitative comparée (ou QCA pour *qualitative comparative analysis*), sont plutôt issues des pratiques de recherche fondamentale, alors que les quatre autres (évaluation basée sur la théorie, évaluation réaliste, analyse de contribution, récolte d'incidences) sont issues du champ de l'évaluation. Les comparaisons de niveau macro (Cf. Chapitre 18) consistent à exploiter les variations et les similitudes entre de grandes entités d'analyse (par exemple, États ou régions) à des fins explicatives : par exemple, pour

expliquer les différences entre de grands modèles de politique sociale, ou l'influence d'une configuration particulière de politique familiale sur le taux d'emploi des femmes. L'analyse qualitative comparée est une méthode mixte qui consiste à traduire des données qualitatives en un format numérique afin d'analyser systématiquement quelles configurations de facteurs produisent un résultat donné (Cf. Chapitre 19). Fondée sur une conception alternative de la causalité de type configurationnel, elle est notamment utile pour comprendre pourquoi une même politique peut entraîner certains changements dans certaines circonstances et non dans d'autres.

Développée en réponse aux limites des approches expérimentales et quasi-expérimentales ne permettant pas de saisir comment une intervention produit ses impacts, l'évaluation basée sur la théorie consiste à ouvrir la « boîte noire » de l'action publique en décomposant les différentes étapes de la chaîne causale liant l'intervention à ses résultats finaux (Cf. Chapitre 20). Les chapitres suivants relèvent globalement de cette famille d'approches en évaluation. L'évaluation réaliste (Cf. Chapitre 21) conçoit les politiques publiques comme des interventions qui produisent leurs effets par le biais de mécanismes qui ne se déclenchent que dans des contextes spécifiques. En mettant au jour des configurations contexte – mécanismes – effets, cette approche permet d'établir pour qui, comment et dans quelles circonstances une intervention fonctionne. Particulièrement adaptée aux interventions complexes, l'analyse de contribution (Cf. Chapitre 22) consiste à formuler progressivement des « hypothèses de contribution », dans un processus impliquant les parties prenantes de la politique, pour ensuite tester ces hypothèses de façon systématique à partir d'une diversité de méthodes. La récolte d'incidence (Cf. Chapitre 23), quant à elle, part d'une appréhension large des « incidences », soit des changements observables, pour ensuite retracer si et comment l'intervention a pu jouer un rôle dans leur production. Enfin, le dernier chapitre est consacré à une démarche innovante en évaluation, prenant appui sur le concept de sécurisation culturelle initialement développé dans les sciences infirmières (Cf. Chapitre 24). La

sécurisation culturelle vise à faire en sorte que l'évaluation se déroule d'une façon dite sûre pour les parties prenantes et notamment les communautés minorisées cibles de l'intervention étudiée, c'est-à-dire que le processus d'évaluation évite de reproduire des mécanismes de domination (agression, déni d'identité...) liés à des inégalités structurelles. Pour cela, diverses techniques participatives sont mobilisées à toutes les étapes de l'évaluation. Ce chapitre est ainsi l'occasion d'insister sur l'importance des dynamiques participatives, également soulignée par plusieurs autres contributions.

Une présentation didactique et illustrée

Pour faciliter la lecture et la comparaison entre méthodes et approches, chaque chapitre est organisé selon un plan commun autour de cinq grandes questions :

- 1) En quoi consiste cette méthode/approche?
- 2) En quoi est-elle utile pour l'évaluation des politiques publiques?
- 3) Un exemple de mobilisation de cette méthode/approche;
- 4) Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode/approche?
- 5) Quels sont les atouts et les limites de cette méthode/approche par rapport à d'autres?

L'ouvrage est directement publié en deux langues (français et anglais) afin de favoriser sa diffusion. Les contributions ont été initialement rédigées dans l'une ou l'autre langue selon la préférence des autrices et auteurs, puis traduites et révisées (lorsque possible) par celles-ci et ceux-ci. Un glossaire bilingue est disponible ci-après pour faciliter le passage d'une langue à l'autre.

Les exemples mobilisés relèvent d'une pluralité de domaines de politiques publiques, étudiés dans une diversité de contexte : retraites en Italie, information météorologique et climatique au Sénégal, salaire minimum dans le New Jersey, accueil dans les services publics en France, développement de l'enfant en Chine, lutte contre le tabagisme chez les jeunes au Royaume-Uni, financement de la santé au Burkina Faso, impact d'une école d'été sur la réussite scolaire aux États-Unis, formation aux compétences non techniques en Belgique, développement de la participation citoyenne pour améliorer les services publics en République Dominicaine, projet de nutrition au Bangladesh, couverture santé universelle dans six pays d'Afrique, etc. Les nombreux exemples présentés au fil des chapitres permettent ainsi d'illustrer la diversité et la vitalité actuelle des pratiques de recherche évaluative.

Loin de prétendre à l'exhaustivité, cette publication constitue une première synthèse de quelques méthodes parmi les plus utilisées. La collection a vocation à être enrichie par le biais de publications au fil de l'eau dans la collection en accès ouvert des fiches méthodologiques du LIEPP⁷.

Bibliographie

Baïz, Adam, et Anne Revillard. 2022. *Comment articuler les méthodes qualitatives et quantitatives pour évaluer l'impact des politiques publiques?* Paris: France Stratégie. <https://www.strategie.gouv.fr/publications/articuler-methodes-qualitatives-quantitatives-evaluer-limpact-politiques-publiques>.

7. Fiches méthodologiques LIEPP : <https://www.sciencespo.fr/liepp/fr/publications.html#Fiches%20m%C3%A9thodologiques%20LIEPP>

- Belaid, Loubna, et Valéry Ridde. 2020. « Une cartographie de quelques méthodes de revues systématiques ». *Working Paper CEPED n°44*.
- Brisolara, Sharon. 1998. « The history of participatory evaluation and current debates in the field ». *New Directions for Evaluation* 1998 (80): 25-41. <https://doi.org/10.1002/ev.1115>.
- Burch, Patricia, et Carolyn J. Heinrich. 2016. *Mixed Methods for Policy Research and Program Evaluation*. Los Angeles: Sage.
- Campbell, Donald T., et Julian C. Stanley. 1963. « Experimental and quasi-experimental designs for research ». In *Handbook of research on teaching*. Houghton Mifflin Company.
- Cousins, J. Bradley, et Elizabeth Whitmore. 1998. « Framing participatory evaluation ». *New Directions for Evaluation*, no 80: 5-23.
- Cox, Gary B. 2021. « La malédiction de l'évaluation au sein des universités ». In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 409-15. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/la-malediction-de-levaluation-au-sein-des-universites/>.
- Delahais, Thomas. 2022. « Le choix des approches évaluatives ». In *L'évaluation en contexte de développement: Enjeux, approches et pratiques*, par Linda Rey, Jean Serge Quesnel, et Vénétia Sauvain, 155-80. Montréal: JFP/ENAP.
- Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, éd. 2021. *Evaluation: Fondements, controverses, perspectives*. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/>.
- Desplat, Rozenn, et Marc Ferracci. 2017. *Comment évaluer l'impact des politiques publiques? Un guide à l'usage des décideurs et praticiens*. Paris: France Stratégie.

- Devaux-Spatarakis, Agathe. 2014. « L'expérimentation « telle qu'elle se fait » : leçons de trois expérimentations par assignation aléatoire ». *Formation emploi*, no 126: 17-38.
- Duran, Patrice, Christine Erhel, et Jérôme Gautié. 2018. « L'évaluation des politiques publiques ». *Idées économiques et sociales* 193 (3): 4-5.
- Duran, Patrice, Eric Monnier, et Andy Smith. 1995. « Evaluation à La Française: Towards a New Relationship between Social Science and Public Action ». *Evaluation* 1 (1): 45-63. <https://doi.org/10.1177/135638909500100104>.
- Fougère, Denis, et Nicolas Jacquemet. 2019. « Causal inference and impact evaluation ». *Economie et Statistique*, no 510-511-512: 181-200. <https://doi.org/10.24187/ecostat.2019.510t.1996>.
- Greene, Jennifer C., Lehn Benjamin, et Leslie Goodyear. 2001. « The merits of mixing methods in evaluation ». *Evaluation* 7 (1): 25-44.
- Hong, Quan Nha, et Pierre Pluye. 2018. « Systematic reviews: A brief historical overview ». *Education for information*, no 34: 261-76.
- Jacob, Steve, Alkin, Marvin, et Christina Christie. 2012. « An Evaluation Theory Tree ». In *Evaluation Roots*, édité par Marvin Alkin et Christina Christie. London: Sage. <https://doi.org/10.4135/9781412984157.n2>.
- Jacob, Steve. 2021. « L'hybridation disciplinaire, nouveau talisman de l'évaluation? » In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 422-41. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/lhybridation-disciplinaire-nouveau-talisman-de-levaluation/>.
- Mathison, Sandra. 2005. *Encyclopedia of evaluation*. London: Sage.

- Maxwell, Joseph A. 2004. « Using Qualitative Methods for Causal Explanation ». *Field Methods* 16 (3): 243-64. <https://doi.org/10.1177/1525822X04266831>.
- . 2012. *A Realist Approach for Qualitative Research*. London: Sage.
- Mayne, John. 2012. « Contribution analysis: Coming of age? » *Evaluation* 18 (3): 270-80. <https://doi.org/10.1177/1356389012451663>.
- Mertens, Donna M. 2017. *Mixed Methods Design in Evaluation*. Vol. 1. Evaluation in Practice Series. Thousand Oaks: SAGE Publications, Incorporated. <https://doi.org/10.4135/9781506330631>.
- Mohr, Lawrence B. 1999. « The Qualitative Method of Impact Analysis ». *American Journal of Evaluation* 20 (1): 69-84. <https://doi.org/10.1177/109821409902000106>.
- . 2021. « La méthode qualitative d'analyse d'impact ». In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 500-507. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/la-methode-qualitative-danalyse-dimpact/>.
- Newcomer, Kathryn E., Harry P. Hatry, et Joseph S. Wholey. 2015. *Handbook of Practical Program Evaluation*. Hoboken: Wiley.
- Patton, Michael Q. 1997. *Utilization-focused evaluation*. 3e éd. Thousand Oaks: Sage Publications.
- . 2015. *Qualitative research and evaluation methods: integrating theory and practice*. London: Sage.
- . 2018. *Utilization-focused evaluation*. London: Sage.
- Pawson, Ray, et Nicholas Tilley. 1997. *Realistic evaluation*. London: Sage.

Revillard, Anne. 2018. *Quelle place pour les méthodes qualitatives dans l'évaluation des politiques publiques?* Paris: LIEPP Working Paper n°81.

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

———. 2020. *Évaluation des interventions de santé mondiale: méthodes avancées*. Québec: Éditions science et bien commun.

Rogers, Patricia J., et Carol H. Weiss. 2007. « Theory-Based Evaluation: Reflections Ten Years on: Theory-Based Evaluation: Past, Present, and Future ». *New Directions for Evaluation* 2007 (114): 63-81. <https://doi.org/10.1002/ev.225>.

Rossi, Peter H., Mark W. Lipsey, et Howard E. Freeman. 2004. *Evaluation: a systematic approach*. London: Sage.

Scriven, Michael. 2021. « De quelques leçons durement acquises en évaluation de programme ». In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 416-21. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/de-quelques-lecons-durement-acquises-en-evaluation-de-programme/>.

Suchman, Edward A. 2021. « La recherche évaluative: principes et pratiques applicables aux services publics et aux programmes sociaux ». In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 388-95. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/la-recherche-evaluative-principes-et-pratiques-applicables-aux-services-publics-et-aux-programmes-sociaux/>.

- Wanzer, Dana. 2021. « Qu'est-ce que l'évaluation? En quoi diffère-t-elle (ou non) de la recherche? » In *Evaluation. Fondements, controverses, perspectives*, par Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, et Ridde, Valéry, 442-47. Québec: Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/chapter/quest-ce-que-levaluation-en-quoi-differe-t-elle-ou-non-de-la-recherche/>.
- Wasmer, Etienne, et Christine Musselin. 2013. *Évaluation des politiques publiques: faut-il de l'interdisciplinarité?* Paris: LIEPP Methodological discussion paper n°2.
- Weiss, Carol H. 1997. « Theory-Based Evaluation: Past, Present, and Future ». *New Directions for Evaluation* 1997 (76): 41-55. <https://doi.org/10.1002/ev.1086>.
- . 1998. *Evaluation: Methods for Studying Programs and Policies*. Upper Saddle River, NJ: Prentice-Hall.
- Wilson-Grau, Ricardo. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.
- Woodward, James. 2003. *Making things happen a theory of causal explanation*. Oxford studies in philosophy of science. New York: Oxford University Press.

Glossaire

Français	Anglais	Français	Anglais
Affectation aléatoire	<i>Random assignment</i>	Evaluation réaliste	<i>Realist evaluation</i>
Amplitude d'intervalle	<i>Interval amplitude</i>	Expérimentation en laboratoire	<i>Laboratory experimentation</i>
Approche narrative	<i>Narrative approach</i>	Fenêtre d'observation	<i>Observation window</i>
Archives ouvertes compliant les résultats d'évaluations déjà réalisées	<i>Evidence repository</i>	Focus group	<i>Focus group</i>
Attitudes	<i>Attitudes</i>	Généralisation analytique	<i>Analytical generalisation</i>
Biais de comportement non conscients	<i>Non-conscious behavioural bias</i>	Groupe de contrôle synthétique/artificiel	<i>Synthetic/artificial control group</i>
Cartographie cognitive floue	<i>Fuzzy cognitive mapping</i>	Groupes expérimentaux/de traitement et de contrôle	<i>Experimental/treatment and control groups</i>
Cas unique/multiples	<i>Single/multiple cases</i>	Hypothèses de contribution	<i>Contributing claims</i>
Causalité asymétrique	<i>Asymmetrical causality</i>	Idéaux-types	<i>Ideal-types</i>
Causalité conjoncturelle	<i>Conjunctural causation</i>	Identification systématique de schémas croisés	<i>Systematic identification of cross patterns</i>
Chaîne causale	<i>Causal chain</i>	Impact	<i>Impact</i>
Changements observables	<i>Observable changes</i>	Induction	<i>Induction</i>
Chemin d'impact	<i>Impact path</i>	Interprétativisme	<i>Interpretivism</i>
Chemins d'impact	<i>Causal pathways</i>	Intervale de confiance	<i>Confidence interval</i>
Chemins de contribution	<i>Contribution pathways</i>	Interventions complexes	<i>Complex interventions</i>
Combinaisons de conditions	<i>Combinations of conditions</i>	Logigramme	<i>Flow chart</i>

Comparaison	<i>Comparison</i>	Logique d'intervention	<i>Intervention logic</i>
Complexité causale	<i>Causal complexity</i>	Méthode expérimentale	<i>Experimental method</i>
Comportements	<i>Behaviours</i>	Méthode mixte	<i>Mixed method</i>
Configuration contexte-mécanisme-effet (cme)	<i>Context-mechanism-outcome (cmo) configuration</i>	Méthode qualitative	<i>Qualitative method</i>
Configurations	<i>Configurations</i>	Méthode quantitative	<i>Quantitative method</i>
Constructivisme	<i>Constructivism</i>	Méthodes quasi-expérimentales	<i>Quasi-experimental methods</i>
Contamination	<i>Contamination</i>	Microsimulation statique/dynamique	<i>Static/dynamic microsimulation</i>
Coût/efficacité	<i>Cost-effectiveness,</i>	Mise en oeuvre des politiques publiques	<i>Policy implementation</i>
Décolonialité	<i>Decoloniality</i>	Modélisation	<i>Modeling</i>
Démarche abductive	<i>Abductive approach</i>	Monotonie	<i>Monotonicity</i>
Dépendance au sentier emprunté	<i>Path dependency</i>	Observation directe	<i>Direct observation</i>
Devis/design	<i>Design</i>	Point d'inflexion	<i>Critical juncture</i>
Devis/design convergent	<i>Convergent design</i>	Preuves	<i>Evidence</i>
Devis/design séquentiel explicatif	<i>Sequential explanatory design</i>	Principes causaux	<i>Causal principles</i>
Devis/design séquentiel exploratoire	<i>Sequential exploratory design</i>	Raisonnement bayésien	<i>Bayesian reasoning</i>
Diagramme logique d'impact	<i>Impact logic diagram</i>	Réalisme critique	<i>Critical realism</i>

Dimension longitudinale des données	<i>Longitudinal dimension of data</i>	Récolte d'incidences	<i>Outcome harvesting</i>
Doubles/triples différences	<i>Difference-in-differences</i>	Régression sur discontinuité stricte/floue	<i>Strict/fuzzy regression discontinuity</i>
Ecart-type	<i>Standard deviation</i>	Response automatique/non-automatique	<i>Automatic/non-automatic response</i>
Efficacité	<i>Effectiveness</i>	Revue de la littérature	<i>Literature review</i>
Efficience	<i>Efficiency</i>	Revue mixte de la littérature	<i>Mixed methods literature review</i>
Empreintes digitales	<i>Fingerprints</i>	Revue systématique mixte	<i>Systematic mixed review</i>
Enoncé d'incidences	<i>Outcome statements</i>	Score de propension	<i>Propensity score</i>
Entretien	<i>Interview</i>	Seuil d'éligibilité	<i>Eligibility threshold</i>
Entretien de groupe	<i>Group interview</i>	Similitudes	<i>Similarities</i>
Entretien semi-directif	<i>Semi-structured interview</i>	Tendances parallèles	<i>Parallel trends</i>
Equifinalité	<i>Equifinality</i>	Théorie de moyenne portée	<i>Middle range theory</i>
Equilibrage par entropie	<i>Entropy balancing</i>	Théorie du changement	<i>Theory of change</i>
Ethnographie	<i>Ethnography</i>	Théorie du changement relative aux processus (tcp)	<i>Process theory of change (ptoc)</i>
Etude de cas	<i>Case study</i>	Traçage de processus	<i>Process tracing</i>
Evaluation attentive aux différences culturelles	<i>Culturally sensitive evaluation</i>	Triangulation empirique	<i>Empirical triangulation</i>
Evaluation autochtone	<i>Indigenous evaluation</i>	Unités macro-sociales	<i>Macro-social units</i>

Evaluation basée sur la théorie	<i>Theory-based evaluation</i>	Validité interne/ externe	<i>Internal/ external validity</i>
Evaluation ex post	Ex post evaluation	Variable de forçage	<i>Forcing variable</i>

Anglais	Français	Anglais	Français
Abductive approach	<i>Démarche abductive</i>	Ideal-types	<i>Idéaux-types</i>
Analytical generalisation	<i>Généralisation analytique</i>	Impact	<i>Impact</i>
Asymmetrical causality	<i>Causalité asymétrique</i>	Impact logic diagram	<i>Diagramme logique d'impact</i>
Attitudes	<i>Attitudes</i>	Impact path	<i>Chemin d'impact</i>
Automatic/non-automatic response	<i>Response automatique/non-automatique</i>	Indigenous evaluation	<i>Evaluation autochtone</i>
Bayesian reasoning	<i>Raisonnement bayésien</i>	Induction	<i>Induction</i>
Behaviours	<i>Comportements</i>	Internal/external validity	<i>Validité interne/externe</i>
Case study	<i>Etude de cas</i>	Interpretivism	<i>Interprétativisme</i>
Causal chain	<i>Chaîne causale</i>	Interval amplitude	<i>Amplitude d'intervalle</i>
Causal complexity	<i>Complexité causale</i>	Intervention logic	<i>Logique d'intervention</i>
Causal pathways	<i>Chemins d'impact</i>	Interview	<i>Entretien</i>
Causal principles	<i>Principes causaux</i>	Laboratory experimentation	<i>Expérimentation en laboratoire</i>
Combinations of conditions	<i>Combinaisons de conditions</i>	Literature review	<i>Revue de la littérature</i>
Comparison	<i>Comparaison</i>	Longitudinal dimension of data	<i>Dimension longitudinale des données</i>
Complex interventions	<i>Interventions complexes</i>	Macro-social units	<i>Unités macro-sociales</i>
Confidence interval	<i>Intervale de confiance</i>	Middle range theory	<i>Théorie de moyenne portée</i>
Configurations	<i>Configurations</i>	Mixed method	<i>Méthode mixte</i>
Conjunctural causation	<i>Causalité conjoncturelle</i>	Mixed methods literature review	<i>Revue mixte de la littérature</i>

Constructivism	<i>Constructivisme</i>	Modeling	<i>Modélisation</i>
Contamination	<i>Contamination</i>	Monotonicity	<i>Monotonïcité</i>
Context-mechanism-outcome (CMO) configuration	<i>Configuration contexte-mécanisme -effet (CME)</i>	Narrative approach	<i>Approche narrative</i>
Contributing claims	<i>Hypothèses de contribution</i>	Non-conscious behavioural bias	<i>Biais de comportement non conscients</i>
Contribution pathways	<i>Chemins de contribution</i>	Observable changes	<i>Changements observables</i>
Convergent design	<i>Devis/design convergent</i>	Observation window	<i>Fenêtre d'observation</i>
Cost-effectiveness,	<i>Coût/efficacité</i>	Outcome harvesting	<i>Récolte d'incidences</i>
Critical juncture	<i>Point d'inflexion</i>	Outcome statements	<i>Enoncé d'incidences</i>
Critical realism	<i>Réalisme critique</i>	Parallel trends	<i>Tendances parallèles</i>
Culturally sensitive evaluation	<i>Évaluation attentive aux différences culturelles</i>	Path dependency	<i>Dépendance au sentier emprunté</i>
Decoloniality	<i>Décolonialité</i>	Policy implementation	<i>Mise en oeuvre des politiques publiques</i>
Design	<i>Devis/design</i>	Process theory of change (ptoc)	<i>Théorie du changement relative aux processus (TCP)</i>
Difference-in-differences	<i>Doubles/triples différences</i>	Process tracing	<i>Traçage de processus</i>
Direct observation	<i>Observation directe</i>	Propensity score	<i>Score de propension</i>
Effectiveness	<i>Efficacité</i>	Qualitative method	<i>Méthode qualitative</i>
Efficiency	<i>Efficience</i>	Quantitative method	<i>Méthode quantitative</i>

Eligibility threshold	<i>Seuil d'éligibilité</i>	Quasi-experimental methods	<i>Méthodes quasi-expérimentales</i>
Empirical triangulation	<i>Triangulation empirique</i>	Random assignment	<i>Affectation aléatoire</i>
Entropy balancing	<i>Équilibrage par entropie</i>	Realist evaluation	<i>Évaluation réaliste</i>
Equifinality	<i>Equifinalité</i>	Semi-structured interview	<i>Entretien semi-directif</i>
Ethnography	<i>Ethnographie</i>	Sequential explanatory design	<i>Devis/design séquentiel explicatif</i>
Evidence	<i>Preuves</i>	Sequential exploratory design	<i>Devis/design séquentiel exploratoire</i>
Evidence repository	<i>Archives ouvertes compliant les résultats d'évaluations déjà réalisées</i>	Similarities	<i>Similitudes</i>
Ex post evaluation	<i>Évaluation ex post</i>	Single/multiple cases	<i>Cas unique/multiples</i>
Experimental method	<i>Méthode expérimentale</i>	Standard deviation	<i>Ecart-type</i>
Experimental/treatment and control groups	<i>Groupes expérimentaux/de traitement et de contrôle</i>	Static/dynamic microsimulation	<i>Microsimulation statique/dynamique</i>
Fingerprints	<i>Empreintes digitales</i>	Strict/fuzzy regression discontinuity	<i>Régression sur discontinuité stricte/floue</i>
Flow chart	<i>Logigramme</i>	Synthetic/artificial control group	<i>Groupe de contrôle synthétique/artificiel</i>
Focus group	<i>Focus group</i>	Systematic identification of cross patterns	<i>Identification systématique de schémas croisés</i>
Forcing variable	<i>Variable de forçage</i>	Systematic mixed review	<i>Revue systématique mixte</i>

Fuzzy cognitive mapping	<i>Cartographie cognitive floue</i>	Theory of change	<i>Théorie du changement</i>
Group interview	<i>Entretien de groupe</i>	Theory-based evaluation	<i>Evaluation basée sur la théorie</i>

PARTIE I

MÉTHODES QUANTITATIVES

I. Essais contrôlés randomisés

CARLO BARONE

Résumé

Les essais contrôlés randomisés (ECR) visent à mesurer l'impact d'une intervention donnée en comparant les résultats d'un groupe expérimental (recevant l'intervention) et d'un groupe de contrôle (ne la recevant pas), auxquels les individus sont assignés de manière aléatoire. Il s'agit d'une méthode quantitative utile d'évaluation *ex ante*, pour tester l'impact d'un programme à un stade où il n'a pas encore atteint la totalité de sa population cible (ce qui rend le groupe de contrôle possible).

Mots-clés : Méthodes quantitatives, méthode expérimentale, groupes expérimentaux/de traitement et de contrôle, affectation aléatoire, traitement, contamination

I. En quoi consiste cette méthode?

Les essais contrôlés randomisés (ECR) évaluent l'impact d'une politique en comparant deux groupes : l'un d'eux se voit accorder l'accès à la politique (groupe expérimental), tandis que l'autre est temporairement exclu de la politique (groupe témoin ou contrôle). L'équipe de recherche traduit les objectifs de la politique en mesures quantitatives de résultats et évalue l'efficacité de la politique en comparant ces résultats dans les deux groupes. Si le groupe expérimental affiche des meilleurs résultats en moyenne sur ces mesures, nous concluons que la politique est efficace. Toutefois, cette conclusion est valable si, et seulement si, nous pouvons supposer que les deux groupes étaient parfaitement équivalents. C'est

pourquoi l'affectation aux deux groupes doit se faire de manière aléatoire : si l'échantillon est suffisamment grand, l'affectation aléatoire garantit que les deux groupes sont, en moyenne, initialement équivalents sur toutes les caractéristiques, connues ou non par l'équipe de recherche, mesurées ou non dans l'étude d'évaluation. Par conséquent, toute différence dans les résultats observés après la mise en œuvre de la politique peut être interprétée comme un impact causal de la politique.

Pour réaliser un ERC, on tire un échantillon d'individus qu'on invite à participer à l'étude, en leur expliquant qu'ils et elles peuvent être affecté·e·s soit au groupe expérimental, soit au groupe témoin. Parmi les participant·e·s qui ont accepté de participer, la moitié sera affectée aléatoirement au groupe expérimental et l'autre moitié au groupe témoin. Ce ratio 50%-50% est le plus courant car il permet d'obtenir des estimations plus précises que les ratios non équilibrés (par exemple 70%-30%). En amont de réaliser l'intervention, il est possible d'effectuer une première mesure des résultats (baseline) qui sont mesurés à l'issue de l'intervention chez les participants après, pour que cela serve de base de référence. Cette mesure n'est pas strictement nécessaire, mais elle est souvent effectuée pour plusieurs raisons, par exemple parce qu'elle permet d'étudier les impacts du traitement de manière plus dynamique en comparant les variations des résultats entre les deux groupes avant et après l'intervention. Cette mesure n'est pas strictement nécessaire, mais elle est souvent effectuée pour plusieurs raisons, par exemple parce qu'elle permet d'étudier les impacts du traitement de manière plus dynamique en comparant les variations des résultats entre les deux groupes.

Si la randomisation est une condition nécessaire pour pouvoir émettre des inférences causales plausibles concernant l'intervention, elle n'est pas une condition suffisante. En particulier, le groupe de contrôle doit rester exclu pendant toute la période de mise en œuvre de la politique, c'est-à-dire que nous devons éviter toute forme de "contamination" du traitement. Cela implique, par exemple, que les individus des deux groupes ne communiquent pas sur les objectifs et les contenus du

traitement. De plus, lorsque les individus sont affecté·e·s au groupe de contrôle, ils et elles peuvent réagir en essayant de remplacer le traitement par un traitement similaire. La contamination et le remplacement du traitement peuvent invalider les inférences causales s'ils se produisent à grande échelle. Par conséquent, la condition essentielle est que le groupe de contrôle agisse « comme d'habitude » et il est important que l'équipe de recherche conçoive et présente l'étude de manière à garantir que ce soit le cas. Ainsi, si la randomisation est importante, il est tout aussi important de garantir le plus haut degré de contrôle de ces conditions expérimentales. L'expression « essai contrôlé randomisé » décrit donc les deux conditions essentielles à la réalisation d'inférences causales solides : la répartition aléatoire et le contrôle des conditions expérimentales.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Les ECR visent à estimer les impacts causaux des politiques, c'est-à-dire à évaluer si les politiques produisent des changements correspondant aux objectifs de ces politiques. Le principal défi est que, même si une politique donnée est totalement inefficace, des changements peuvent intervenir en raison d'autres politiques, ou d'autres paramètres économiques ou socioculturels. Par exemple, on peut proposer un programme de formation à des chômeurs et chômeuses pour améliorer leur employabilité et observer ensuite les taux d'emploi des personnes participant à ce programme. Cependant, il n'est pas certain que le changement observé dans ce résultat puisse être attribué à la politique. Par exemple, il pourrait être dû au cycle économique ainsi qu'à tout autre type de politique économique, de travail ou de protection sociale (par exemple, des incitations fiscales à l'embauche, des modifications des règles d'éligibilité aux allocations de chômage, etc.). Par conséquent, une simple comparaison avant-après sans groupe témoin de comparaison ne permet pas d'isoler le véritable impact causal de cette politique.

Les ECR ne sont pas le seul type de méthode d'évaluation de l'impact causal, par exemple les modèles de régression avec discontinuité sont une autre option (voir chapitre séparé). Les ECR sont une forme d'évaluation *ex ante*, c'est-à-dire qu'ils doivent être réalisés avant que la politique ne soit appliquée à l'ensemble de la population des bénéficiaires potentiels. En effet, les ECR supposent que la politique ne soit pas appliquée à certains individus, qui constituent le groupe de contrôle. Si la politique a déjà été généralisée, les ECR sont irréalisables. On peut alors recourir à d'autres types de méthodes d'évaluation de l'impact causal pour isoler le véritable impact causal de la politique.

III. Un exemple d'application : quels sont les messages qui favorisent le mieux la conformité fiscale?

La conformité fiscale (*tax compliance*), c'est-à-dire la déclaration véridique des revenus imposables et le paiement des impôts en temps voulu, est essentielle pour financer les services publics. Une équipe de recherche s'est associée à l'administration fiscale belge pour tester l'impact de différents messages encourageant la conformité fiscale (De Neve et al, 2019). Entre 2014 et 2016, l'équipe a assigné de manière aléatoire environ 2,5 millions de contribuables à recevoir différents messages : des messages simplifiés présentant les informations clés en termes plus simples, des messages de dissuasion visant à rendre explicites les conséquences de la non-conformité, et des messages de morale fiscale visant à motiver les contribuables à apprécier l'importance de la conformité pour la fourniture de biens publics. Les 4 millions de contribuables restant-e-s ont été affecté-e-s à un groupe témoin où la communication avec les contribuables est restée inchangée (cette taille d'échantillon est exceptionnelle, la plupart des ECR se basant sur quelques centaines ou milliers de cas). À l'aide de données administratives, l'équipe de recherche a mesuré l'impact de l'intervention sur la probabilité d'effectuer un paiement ou de déclarer ses impôts,

ainsi que sur le montant des revenus déclarés. Une communication plus simple a eu l'effet le plus important sur le respect des obligations fiscales, incitant les gens à déclarer et à payer leurs impôts plus tôt. L'ajout de messages de dissuasion a encore amélioré le respect des obligations fiscales, tandis que les messages moralisateurs se sont révélés inefficaces.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Dans certains contextes, les ERC sont irréalisables car les risques de contamination ou de remplacement du traitement sont trop élevés. Par exemple, lorsque les individus peuvent facilement communiquer sur le contenu informatif d'une intervention et sont très motivé·e·s pour le faire. Certaines politiques ne peuvent pas être testées avec un ECR car, par construction, elles impliquent l'ensemble de la population, nous ne pouvons donc pas exclure temporairement le groupe de contrôle. C'est par exemple le cas de plusieurs politiques macroéconomiques ou de défense (par exemple, un changement dans les dépenses militaires).

De plus, alors qu'on assigne le plus souvent des individus au groupe de traitement ou de contrôle, on peut aussi assigner des familles entières, des rues ou des villages au traitement ou au contrôle. C'est le cas, par exemple, lorsqu'une intervention donnée est plus efficace, ou ne peut être mise en œuvre, qu'à ces niveaux supra-individuels. Ces types de randomisations de niveau supérieur (randomisation en grappes) peuvent être nécessaires ou extrêmement pratiques, mais ils exigent des échantillons de grande taille et donc des budgets importants.

Enfin, il faut garder à l'esprit que la validité interne (c'est-à-dire la force des inférences causales dans le cas étudié) n'est qu'un des critères de qualité de la recherche en évaluation. Un autre critère important est la validité externe, c'est-à-dire la possibilité de généraliser des conclusions

au-delà de l'échantillon étudié. Ce deuxième critère, lorsqu'il est appliqué aux ECR, exige des échantillons importants et aléatoires de la population étudiée, et que le nombre de participant-e-s abandonnant l'étude reste limité. Un troisième critère important concerne la validité et la fiabilité des mesures des résultats, y compris la capacité d'observer les résultats à long terme d'une politique, et la couverture de tous les effets potentiels (positifs et négatifs) de la politique.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Comme expliqué ci-dessus, la principale force des ECR est qu'ils permettent d'évaluer l'impact causal réel d'une politique avant de l'appliquer à l'ensemble de la population des bénéficiaires. Dans la recherche clinique, les ECR sont la méthode standard pour évaluer l'efficacité de tout type de thérapie ou de médicament et ils sont de plus en plus utilisés pour l'évaluation des politiques publiques, notamment dans les domaines de l'éducation, du marché du travail, de la santé et du logement.

Les applications les plus courantes de cette méthode impliquent la randomisation entre deux groupes d'individus. Cependant, il arrive que l'on puisse organiser trois groupes d'individus ou plus afin de comparer des variantes qualitativement différentes d'une intervention ou des dosages différents de l'intervention. Par exemple, dans une étude visant à promouvoir l'utilisation des services de vélo en libre-service, on peut comparer le groupe de contrôle à un premier groupe de traitement disposant d'informations sur le vélo en libre-service, à un deuxième groupe de traitement recevant une incitation financière à son utilisation et à un troisième groupe recevant une incitation financière plus importante.

Les ECR ne sont pas toujours facilement réalisables. En particulier, les responsables politiques ou les participant·e·s potentiel·le·s peuvent refuser le principe de la randomisation, par exemple parce qu'ils pensent que cela pose un problème éthique car elles excluent les individus du groupe de contrôle des avantages de la politique. Cette critique oublie toutefois que l'exclusion est temporaire, c'est-à-dire qu'elle ne dure que le temps nécessaire pour démontrer que la politique est efficace. Cette exclusion temporaire permet d'évaluer si la politique est efficace avant de la généraliser à l'ensemble de la population. En outre, les ressources disponibles dans les études d'évaluation *ex ante* ne permettent de traiter qu'une petite partie de la population totale, de sorte qu'il serait de toute façon impossible de traiter tout le monde : l'assignation aléatoire donne à chacun les mêmes chances d'être traité.

Il est primordial que les équipes de recherche expliquent en termes simples pourquoi la randomisation est éthique et pourquoi elle est nécessaire pour garantir la fiabilité des comparaisons entre les deux groupes. Lorsque cela est possible, l'acceptabilité sociale de la randomisation peut être accrue en créant une liste d'attente, c'est-à-dire que le groupe témoin reçoit la politique à la fin de l'étude, ou un traitement compensatoire (un traitement différent de celui étudié et qui n'affecte pas le résultat de l'étude). Par exemple, dans une étude fournissant des informations sur les services de garde d'enfants aux mères enceintes afin d'améliorer le recours à ces services, le groupe de contrôle peut recevoir cette information à la fin de l'étude ou recevoir un autre type d'information, par exemple sur les pratiques saines pendant la grossesse. Toutefois, si une liste d'attente est créée, il n'est pas possible d'observer les résultats à long terme car le groupe de contrôle n'est plus exclu de l'intervention. Les listes d'attente et les traitements compensatoires peuvent également être utilisés pour réduire le risque que les personnes assignées au groupe de contrôle abandonnent le traitement. Il est en effet important que les taux d'abandon des deux groupes soient similaires afin de préserver leur équivalence tout au long de l'étude.

Par rapport aux expériences en laboratoire, les ECR ont une validité écologique plus élevée, dès lors que l'on étudie des personnes dans des situations de vie réelles et dans des contextes naturels. Par conséquent, le risque que leur comportement soit influencé par la conscience de faire partie d'une étude est moins important. En même temps, par rapport aux expériences en laboratoire, les ECR permettent un degré de contrôle plus faible sur le comportement des participant·e·s. Dans les expériences cliniques et psychologiques, la conscience d'être traité·e est souvent neutralisée par l'administration de placebos au groupe de contrôle, c'est-à-dire des traitements spécifiquement conçus pour n'avoir aucun effet. Dans les politiques sociales, cette pratique est moins courante car nous avons tendance à considérer les avantages découlant de la conscience d'être traité comme faisant partie intégrante de la politique.

Plus fondamentalement, si les ECR sont un outil fiable pour évaluer les impacts causaux des politiques, ils ne sont pas en position de force pour étudier les processus sous-jacents. Par exemple, si un ECR conclut qu'une politique est inefficace ou moins efficace que prévu, cette méthode est incapable d'expliquer ce qui n'a pas fonctionné et comment nous pouvons améliorer cette politique. C'est pourquoi il est extrêmement utile d'intégrer les ECR aux techniques qualitatives d'évaluation des processus. De plus, les croyances et les perceptions de la politique qu'ont les bénéficiaires et les agent·e·s de la mise en oeuvre peuvent être étudiées en utilisant des entretiens qualitatifs ou des enquêtes.

Quelques références bibliographiques pour aller plus loin

De Neve, Jan-Emmanuel. et Imbert, Clement. et Spinnewijn, Johannes. et Tsankova, Teodora. et Luts, Maarten. 2019. "How to Improve Tax Compliance? Evidence from Population-wide Experiments in Belgium." Working paper.

- Gertler, Paul. et Martinez, Sebastian. et Premand, Patrick. et Rawlings, Laura. et Vermeersch, Christel. 2016. *Impact Evaluation in Practice*, deuxième édition. World Bank Group (Chapitres 3 et 4 de ce manuel). <https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464807794.pdf?sequence=2&isAllowed=y>
- Gibson, Michael. et Sautmann, Anja. Dernière mise à jour : avril 2021. *Introduction to randomized evaluations*, Abdul Latif Jameel Poverty Action Lab. <https://www.povertyactionlab.org/resource/introduction-randomized-evaluations>
- White, Howard. et Sabarwal, Shagun. et De Hoop, Thomas. 2014. *Essais Contrôlés Randomisés, Notes méthodologiques, Évaluation d'impact no 7*, Unicef. <https://www.unicef-irc.org/publications/pdf/MB7FR.pdf>

2. Méthode des doubles différences (difference-in-differences)

DENIS FOUGÈRE ET NICOLAS JACQUEMET

Résumé

La méthode des doubles différences est une méthode quantitative quasi-expérimentale permettant d'évaluer l'impact d'une intervention grâce à la constitution de groupes de comparaison et à la mesure de l'évolution d'un résultat entre un moment initial pré-intervention et un moment ultérieur où seulement un des deux groupes a reçu l'intervention. Cette méthode est très utile pour l'évaluation *ex post* de l'impact d'une intervention.

Mots-clés : Méthodes quantitatives, méthodes quasi-expérimentales, doubles/triples différences, dimension longitudinale des données, tendances parallèles, équilibrage par entropie, groupe de contrôle synthétique/artificiel

I. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Bien que l'évaluation des politiques publiques recouvre un ensemble de problématiques et d'outils très large, qui va bien au-delà de la seule quantification de leurs effets, la question de l'efficacité des politiques

mises en place dans le passé est évidemment primordiale, car elle constitue un guide utile pour envisager leur pérennisation, leur évolution, leur généralisation, voire leur abandon.

Une telle évaluation nécessite de définir clairement les objectifs poursuivis. S'ils sont mesurables, l'évaluation revient à essayer d'observer quels ont été les effets de l'intervention sur les différents types d'agents (ménages, chômeuses ou chômeurs, entreprises, régions, etc.) qui en ont bénéficié : par exemple, l'effet sur l'intensité des départs en retraite de l'élévation de l'âge minimum de la retraite, l'impact sur les transitions vers l'enseignement supérieur de la mise en place d'un système de bourses universitaires, ou encore les conséquences sur le recours au système de santé de l'introduction d'aides financières pour faciliter l'accès aux soins. Un réflexe naturel, qui apparaît (trop) souvent dans le débat public, consiste à comparer la situation des personnes qui ont bénéficié des interventions mises en place à celle d'autres personnes qui n'en ont pas bénéficié. Pour évaluer l'efficacité d'une réforme de l'assurance chômage offrant une aide personnalisée à la recherche d'emploi, on pourrait ainsi comparer les personnes qui ont bénéficié de cette aide à celles qui n'ont bénéficié d'aucun accompagnement. Comme l'illustre l'étude de Fougère, Kamionka et Prieto (2010, voir en particulier la Figure 3), une telle comparaison montre sans aucune ambiguïté que les programmes d'aide à la recherche d'emploi conduisent à un retour à l'emploi beaucoup plus lent pour les personnes qui en ont bénéficié. Faut-il en conclure que les services proposés nuisent à la capacité des personnes en recherche d'emploi à s'insérer sur le marché du travail?

Bien évidemment non. Il faut en conclure, selon le vieil adage, que comparaison n'est pas raison, et que les personnes à qui l'on propose un accompagnement sont précisément celles qui éprouvent les difficultés les plus grandes à trouver un emploi. Lorsque l'on compare leur situation à celle des chômeuses et chômeurs qui n'ont pas bénéficié d'une aide, on fait l'hypothèse implicite que le retour à l'emploi observé dans cette catégorie peut servir de point de référence (de « *contrefactuel* ») à la situation qu'auraient connu les bénéficiaires en l'absence

d'accompagnement. Or les bénéficiaires le sont précisément parce que leur situation eût été particulièrement difficile en l'absence d'accompagnement. Pour éviter de telles confusions, la méthode des doubles différences consiste à définir les groupes de comparaison de sorte que l'écart observé fournisse une mesure plus convaincante de l'effet de l'intervention.

II. En quoi consiste cette méthode?

Supposons que l'on observe les variations entre deux dates d'une variable de résultat (également appelée variable-réponse ou variable dépendante) au sein de deux groupes distincts. Le premier de ces groupes, appelé groupe cible ou groupe traité, bénéficie d'une intervention ou d'une politique donnée (désignée comme le traitement); le second, appelé groupe témoin ou groupe de contrôle, n'en bénéficie pas. La politique est mise en place entre les deux dates considérées. La mesure de l'effet de l'intervention repose exclusivement sur la variation au cours du temps de la variable de résultat entre ces deux dates. Cette variation diffère dans les deux groupes, en général à partir du moment où le traitement entre en vigueur. C'est cette inflexion dans l'écart entre les deux groupes que l'on interprète ici comme l'effet moyen de la politique sur la variable de résultat.

Pourquoi appelle-t-on ce procédé « méthode des doubles différences »? La première différence correspond à la différence entre la valeur moyenne de la variable de résultat dans le groupe de traitement à la seconde date (après mise en œuvre de la politique à évaluer) et la valeur moyenne de cette même variable dans le même groupe à la date initiale (avant mise en œuvre de la politique à évaluer). À cette première différence, on soustrait ensuite la différence analogue pour le groupe de contrôle. La méthode des doubles différences exploite donc la dimension longitudinale des données (ou pseudo-longitudinale, car les individus qui

appartiennent à chacun des groupes peuvent ne pas rester les mêmes au cours du temps) afin de fournir une évaluation *ex post* des politiques publiques mises en œuvre.

La capacité de cette méthode à mesurer l'effet moyen de l'intervention ne repose pas sur l'hypothèse selon laquelle les non-bénéficiaires peuvent servir de groupe de référence aux bénéficiaires en l'absence d'intervention, mais uniquement sur le fait qu'en l'absence d'intervention, l'évolution moyenne de la variable de résultat pour les individus du groupe traité aurait été la même que celle observée au sein du groupe de contrôle (hypothèse de tendances parallèles, « *parallel trends* »). La validité de cette hypothèse, non vérifiable, peut être confortée par le fait qu'avant la mise en place de la politique, la variable de résultat a évolué de la même façon dans les deux groupes (hypothèse de tendances communes, « *common pre-trends* »). À l'inverse de la précédente, cette seconde hypothèse peut être testée grâce aux données observées préalablement à la mise en place de l'intervention, à condition de disposer d'une profondeur d'observation suffisante au cours de cette période – par exemple d'au moins cinq observations dans les deux groupes avant la mise en œuvre de la politique évaluée (ces observations sont appelées « *leads* » dans la littérature académique). L'hypothèse de tendances parallèles équivaut à supposer que l'écart préexistant entre les deux groupes, qui provient des divers facteurs conduisant à des niveaux différents de la variable de résultat au sein de ces groupes, serait resté le même en l'absence d'intervention, de sorte que l'évolution constatée de cet écart peut être interprétée comme l'effet moyen de l'intervention.

Cette approche n'est donc valide qu'à condition que l'intervention laisse inchangée la variable de résultat dans le groupe de contrôle (hypothèse appelée SUTVA, « *Stable Unit Treatment Value Assumption* »). En effet, tout effet indirect de l'intervention sur ce groupe (si, par exemple, la difficulté à trouver un emploi augmente parce que l'accélération du retour à l'emploi dans le groupe de traitement accroît la tension sur le marché du travail) remet en cause l'hypothèse de tendances parallèles. De la même façon, cette hypothèse pourrait être remise en cause si le groupe

de traitement anticipait la mise en place de l'intervention (ce qui aurait pour conséquence, par exemple, de ralentir sa recherche d'emploi en raison de la perspective des mesures d'accompagnement à venir). Cette violation de l'hypothèse de tendances parallèles est connue sous le terme d'« *Ashenfelter gap* ».

Compte tenu des nombreux facteurs susceptibles d'affecter la validité de l'approche, les développements récents de la méthode des doubles différences visent notamment à affiner la constitution des groupes de manière à augmenter leur comparabilité (voir Roth et alii, 2022, pour une description détaillée). Il est ainsi possible de recourir aux méthodes de « *matching* » (voir chapitre séparé) qui associent, à l'aide d'un critère statistique, à chaque personne bénéficiaire de l'intervention la ou les personnes du groupe de contrôle dont les caractéristiques observables sont proches – de sorte que la comparaison est réalisée entre « *plus proche voisins* » statistiques – ou bien encore à la méthode de l'équilibrage par entropie (*entropy balancing*) qui permet d'égaliser les premiers moments (moyenne, variance, coefficient d'asymétrie, etc.) des distributions des variables explicatives observables au sein des deux groupes. Une démarche similaire peut-être appliquée à la variable de résultat elle-même plutôt qu'à la distribution des caractéristiques observables. C'est l'objectif du *groupe de contrôle synthétique*, qui consiste à créer par un système adéquat de pondérations un *groupe de contrôle artificiel* à partir des observations du groupe de contrôle. Ce groupe de contrôle synthétique est construit de telle sorte que l'évolution passée de la variable de résultat en son sein soit identique à celle de la même variable dans le groupe de traitement en minimisant, à l'aide d'un système de pondération des observations du groupe de contrôle, la distance relative à la variable de résultat entre le groupe traitement avant intervention et ce groupe de contrôle synthétique. Dans le cas où le nombre d'unités traitées est très grand, il est possible que le contrôle synthétique d'une unité traitée ne soit pas unique. Plusieurs contributions récentes ont proposé des solutions permettant de résoudre cette

difficulté. Parmi celles-ci, certaines suggèrent d'utiliser des techniques de complétion de matrices, d'autres proposent des méthodes d'inférence fondées sur les échantillons (*sampling-based inferential methods*).

L'une des extensions les plus populaires destinées à prendre en compte l'existence d'interactions non observables entre les caractéristiques de groupe et de temps que la méthode des doubles différences pourrait omettre est la méthode des « triples différences » (*difference in difference in differences*). Cette méthode repose sur l'observation de deux groupes supplémentaires, un « faux » groupe de traitement et un « faux » groupe de contrôle. Imaginons par exemple une politique de santé qui s'applique dans une région A aux personnes de plus de 65 ans. Pour évaluer les effets de cette politique sur le recours aux soins et sur l'état de santé des personnes concernées, il est possible de considérer comme groupe de traitement les personnes âgées de 65 à 69 ans de la région A, et utiliser la situation de celles qui sont âgées de 60 à 64 ans dans cette même région comme groupe de contrôle. Une première double différence appliquée à ces deux groupes doit en principe produire une estimation de l'effet moyen de l'intervention sur le recours aux soins et l'état de santé des personnes de plus de 65 ans dans la région A. Mais on peut reprocher à cette approche qu'elle compare des populations qui ne sont pas tout à fait les mêmes du point de vue de leur état de santé : les personnes de 68 ou 69 ans sont probablement en moins bonne santé que celles âgées de 60 ou 61 ans, et donc exposées à des risques de dégradation de leur santé au cours du temps qui sont plus élevés. Pour répondre à cette critique, il est possible de considérer les mêmes groupes d'âge dans une seconde région, la région B, dans laquelle la même politique n'est pas mise en œuvre, puis de calculer un second estimateur des doubles différences dans cette région B. On peut ensuite soustraire cette seconde double différence dans la région B à celle calculée dans la région A. La seconde double différence appliquée aux deux groupes de personnes de la région B élimine les écarts de santé entre groupes d'âge qui prévalent

naturellement dans l'ensemble de la population (l'hypothèse de tendances parallèles est donc affaiblie, et porte ici sur la différence relative entre les deux catégories de population dans chacune des deux régions).

Outre la qualité de la comparaison entre les groupes, une seconde limite de la méthode des doubles différences est que l'effet de l'intervention n'est pas toujours identique au sein de différents sous-groupes de bénéficiaires, ou au cours du temps (l'effet de l'intervention est dit « hétérogène »). En s'appuyant sur l'évolution de l'écart entre deux groupes seulement, cette méthode mesure uniquement un effet moyen, qui n'est compatible qu'avec de très fortes variations de l'effet de l'intervention entre différents sous-groupes. Afin d'étudier les variations de l'effet au cours du temps, il est utile de disposer d'observations de la variable de résultat dans les deux groupes bien au-delà de la date qui suit la mise en œuvre de cette intervention (observations qui sont parfois appelées « *lags* »). Il est ainsi possible de s'assurer que la politique évaluée a bien des effets significatifs à moyen terme, voire à long terme si le suivi statistique est suffisamment long.

Une telle hétérogénéité de l'effet de l'intervention soulève également des difficultés importantes lorsque sa diffusion dans le groupe de traitement est graduelle. La méthode habituelle, qui consiste à intégrer les observations au groupe des bénéficiaires au fur et à mesure de leur éligibilité à l'intervention conduit en effet à des conclusions infondées (qui peuvent aller jusqu'à conclure à l'inefficacité d'une intervention dont les effets sont positifs pour l'ensemble des bénéficiaires). Les études récentes recommandent de se concentrer uniquement sur les observations qui correspondent à un changement de situation, ce qui revient à combiner de multiples estimateurs des doubles différences calculés à toutes les dates auxquelles le périmètre du groupe de bénéficiaires évolue (voir de Chaisemartin et d'Haultfoeuille, 2022, pour une présentation complète).

III. Un exemple d'utilisation de cette méthode dans le domaine de l'emploi

Comme la plupart des instruments d'intervention sur le marché du travail, l'introduction d'un salaire minimum et la fixation de son niveau résultent d'un arbitrage délicat : sur un marché du travail où les employeurs ont un pouvoir de négociation élevé qui leur permet de comprimer les salaires, le salaire minimum constitue une protection pour les salarié·e·s et permet de répartir les bénéfices de la production de manière plus équitable. Mais l'existence d'un salaire minimum implique également que toutes les activités dont la rentabilité est inférieure au salaire minimum ne pourront pas avoir lieu car elles ne permettent pas de créer une valeur suffisante pour couvrir le coût des salaires. Toute la difficulté est donc de fixer un salaire minimum qui rééquilibre la négociation salariale sans nuire de manière excessive à l'efficacité économique.

L'article de Card et Krueger (1994) sur l'augmentation du salaire minimum décrétée dans le New Jersey en avril 1992, est l'une des études les plus célèbres de la mise en œuvre de la méthode des doubles différences. Dans cette étude, Card et Krueger comparent le niveau d'emploi dans le secteur de la restauration rapide (très intensive en emploi peu qualifié qui est en général rémunéré au salaire minimum) dans le New Jersey et en Pennsylvanie, en février 1992 et en novembre 1992. Ces dates encadrent une augmentation du salaire minimum horaire de 4,25 US dollars à 5,05 US dollars intervenue en avril 1992 dans le New Jersey, alors qu'à la même date, ce salaire restait constant et égal à 4,25 US dollars en Pennsylvanie. Observer une évolution de l'emploi dans le New Jersey entre février et novembre 1992 au moyen d'une première différence ne permet pas d'attribuer cette évolution à la seule hausse du salaire minimum dans cet état, notamment parce que d'autres facteurs concomitants, tels que les conditions météorologiques ou macroéconomiques, pourraient également y contribuer. Par ailleurs, l'écart dans les niveaux d'emploi entre les deux états postérieurement à l'élévation du salaire minimum

reflète non seulement l'effet de cette politique mais aussi l'ensemble des différences de fonctionnement de ce secteur d'activité entre le New Jersey et la Pennsylvanie.

En incluant à la fois les *fastfoods* du New Jersey (le groupe de traitement) et de Pennsylvanie (le groupe de contrôle), situés des deux côtés de la frontière de ces états, Card et Krueger peuvent limiter au moyen d'une seconde différence les effets de ces deux types de facteurs. Sous l'hypothèse de tendances parallèles, l'évolution de l'emploi dans le secteur de la restauration rapide en Pennsylvanie peut être interprétée comme l'évolution de l'emploi qu'aurait connu le secteur de la restauration rapide dans le New Jersey si le salaire minimum horaire n'avait pas augmenté dans cet état. Les estimations réalisées par Card et Krueger suggèrent que l'augmentation du salaire minimum ne s'est pas accompagnée d'une diminution de l'emploi dans le New Jersey. Plus précisément, Card et Krueger estiment que l'augmentation de 0,80 US dollars du salaire minimum horaire dans le New Jersey a entraîné (a « causé ») une augmentation de 2,75 emplois à temps plein en moyenne dans chaque fast-food de cet état.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

L'estimateur obtenu sera d'autant plus informatif (et l'hypothèse de tendances parallèles sera d'autant plus crédible) que le groupe de contrôle est semblable au groupe de traitement du point de vue des caractéristiques explicatives observables (en évitant de surinterpréter de telles comparaisons, puisque l'hétérogénéité inobservable peut varier considérablement entre les groupes sans que cela soit détectable). À moins que la constitution des groupes ne suive une procédure qui impose une telle condition, il convient pour s'en assurer de mettre en regard la distribution des caractéristiques observables entre les groupes (par

exemple, dans un échantillon de salarié-e-s, la proportion de femmes, les différents groupes d'âge et de niveaux d'éducation, etc.) puis de réaliser un ensemble de tests statistiques permettant de vérifier l'absence de différences significatives entre les groupes (la procédure est connue sous le nom de « *balancing test* »). Une bonne pratique consiste à conditionner l'analyse statistique à toute caractéristique observable dont la distribution varie entre les groupes afin de prendre en compte de possibles interactions entre cette caractéristique et les variations au cours du temps.

Afin de vérifier la robustesse des résultats, il est possible de recourir à des groupes dits « placebos » de manière à répliquer l'analyse sur un groupe d'observations n'ayant pas été exposé à l'intervention évaluée. Une première façon de procéder est d'utiliser un « faux » groupe de traitement, qui peut être le même groupe de traitement mais observé à au moins deux dates antérieures à la mise en place de la politique publique évaluée, ou bien encore un troisième groupe qui est supposé ne pas être concerné par la politique mise en œuvre. La robustesse de l'analyse est confortée si cette procédure permet de conclure à l'absence d'effet. Une seconde pratique consiste à utiliser un autre groupe de contrôle, semblable au premier groupe de contrôle utilisé. En ce cas, l'estimation de l'effet moyen du traitement doit être approximativement égale à celle obtenue avec le groupe de contrôle initial.

Si l'utilisation de données répétées dans le temps permet d'améliorer la qualité des comparaisons qui sont faites, elles conduisent à travailler avec des observations qui sont liées entre elles au cours du temps. Cette propriété des données a longtemps été négligée dans l'application de la méthode des doubles différences, ceci conduisant à des mesures de la significativité statistique des effets observés qui sont erronées. Il est donc important de prendre en compte la structure de corrélation des données dans l'analyse statistique (Bertrand et alii, 2004).

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

La méthode des doubles différences est une méthode quasi-expérimentale, dans le sens où elle est principalement utilisée pour étudier des changements qui surviennent spontanément et selon des modalités qui ne sont pas directement liées à l'objectif d'évaluation, mais qui produisent des observations qui permettent de se rapprocher d'une situation expérimentale. Comme toutes les méthodes quasi-expérimentales, les effets qu'elle mesure correspondent aux effets de la politique sur la sous-population qui a effectivement été ciblée et a de fait bénéficié de la politique (dans les termes du modèle d'évaluation causale des politiques publiques, elle mesure un effet de traitement sur les traités, ATT « *Average Treatment on the Treated* »). Dès lors que l'intervention a été ciblée à dessein sur des catégories de population particulières (qui sont particulièrement sensibles à l'intervention mise en œuvre, ou qui en ont particulièrement besoin) cette approche ne permet pas de mesurer l'effet moyen de la politique (ATE, « *Average Treatment Effect* »), c'est-à-dire l'effet qu'elle produirait si elle était généralisée à l'ensemble de la population, ni même les variations de l'effet entre différents individus traités. Athey et Imbens (2006) proposent une approche alternative à la méthode des doubles différences qui fournit une estimation de la totalité de la distribution contrefactuelle de la variable de résultat et permet de mesurer plus finement les variations de l'effet de l'intervention entre différents types de bénéficiaires.

Il n'en reste pas moins que cette méthode permet de mesurer un effet moyen dans une sous-population plus large que la plupart des méthodes quasi-expérimentales existantes. À ce titre, elle se distingue en particulier de la régression sur discontinuité (voir fiche séparée) et de l'estimation de l'effet local moyen du traitement (*local average treatment effect*, ou LATE) qui permettent seulement d'estimer les effets moyens du traitement pour une sous-population particulière, à savoir pour le sous-groupe de

personnes (appelées « compliers » en anglais) dont l'accès au traitement est uniquement dû à leur proximité d'un seuil fixé de manière exogène (par exemple, un seuil d'âge ou de revenu) dans le premier cas, et celles et ceux qui en bénéficient en raison de la variable d'instrumentation dans le second.

Quelques références bibliographiques pour aller plus loin

Athey, Susan. et Imbens, Guido. W.. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models", *Econometrica*, 74(2): 431-97. <https://doi.org/10.1111/j.1468-0262.2006.00668.x>

Bertrand, Marianne. et Duflo, Esther. et Mullainathan Sendhil. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics*, 119(1): 249-275. <https://doi.org/10.1162/003355304772839588>

Card, David. et Krueger, Alan B.. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", *American Economic Review*, 84(4): 772-793. <https://www.jstor.org/stable/2118030>

De Chaisemartin, Clément. et D'Haultfoeuille, Xavier. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects", NBER *Working Paper* No. 29873. DOI 10.3386/w29873

Fougère, Denis. et Kamionka, Thierry. et Prieto, Ana. 2010. « L'efficacité des mesures d'accompagnement sur le retour à l'emploi », *Revue Economique*, 61(3): 599-612. <http://dx.doi.org/10.3917/reco.613.0599>

Roth, Jonathan. et Sant'Anna, Pedro H. C.. et Bilinski, Alyssa. et Poe, John. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature", arXiv:2201.01194, <https://doi.org/10.48550/arXiv.2201.01194>

Des ressources pour mettre en œuvre cette méthode avec les logiciels Stata et R

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press: New Haven and London. Disponible en libre accès sur le site <https://mixtape.scunning.com/index.html>

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*, Chapitre 18. Chapman and Hall/CRC Press: Boca Raton, Florida. Disponible en libre accès sur le site <https://theeffectbook.net/ch-DifferenceinDifference.html>

3. La régression sur discontinuité

DENIS FOUGÈRE ET NICOLAS JACQUEMET

Résumé

La régression sur discontinuité est une méthode quantitative quasi-expérimentale qui évalue l'impact d'une intervention en comparant les observations qui se situent à proximité d'un seuil d'éligibilité fixé par les autorités en charge de la politique étudiée. L'existence d'un tel seuil (par exemple, l'éligibilité à la politique à partir de tel âge, ou en dessous de tel niveau de revenu) permet en effet de reconstituer un groupe cible et un groupe de contrôle, de façon analogue à une démarche expérimentale.

Mots-clés : Méthodes quantitatives, méthodes quasi-expérimentales, seuil d'éligibilité, variable de forçage, régression sur discontinuité stricte/floue, fenêtre d'observation, amplitude d'intervalle, monotonie, personnes observantes

I. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Lorsque l'on souhaite réaliser l'évaluation quantitative des effets d'une politique publique, la difficulté principale consiste à trouver un groupe de comparaison (appelé groupe de contrôle) dont la situation puisse servir de point de référence (i.e., « de contrefactuel » : voir la fiche 'Méthode des doubles différences') à celle que connaissent les bénéficiaires de

l'intervention (groupe cible ou groupe de traitement). L'expérimentation randomisée, dans laquelle les bénéficiaires et les non-bénéficiaires sont choisis aléatoirement dans une population éligible donnée, constitue le cadre de référence idéal pour définir un groupe de contrôle valide : par construction, si l'on dispose d'un échantillon assez grand, le groupe de contrôle aura les mêmes caractéristiques (de genre, d'âge, de niveau d'éducation, etc.) que le groupe de traitement.

Les méthodes quasi-expérimentales visent à pallier l'absence d'expériences contrôlées en s'appuyant sur des variations qui surviennent spontanément (en général, par décision des pouvoirs publics) et produisent des observations qui se rapprochent d'une situation expérimentale. Les méthodes d'appariement ou d'estimation par doubles différences exploitent les cas dans lesquels la mise en place d'une politique publique produit naturellement deux groupes dont la comparaison permet, sous certaines conditions, d'en mesurer l'effet. La méthode de régression sur discontinuité, quant à elle, exploite l'application d'un seuil d'éligibilité pour décider du déclenchement de l'intervention, qui produit l'équivalent d'une expérience randomisée locale au voisinage du seuil.

II. En quoi consiste cette méthode?

Lorsque l'accès à une intervention ou à une politique publique est conditionné par un seuil fixé par les autorités en charge de cette politique, l'intervention produit deux groupes dont l'un seulement bénéficie de l'intervention. Mais ces groupes ne sont pas directement comparables puisqu'ils diffèrent par construction en raison de la valeur de la caractéristique définissant le seuil. Ce seuil peut être une condition d'âge (pour un départ en retraite par exemple), de taille pour les entreprises (par exemple, une politique de réduction des charges destinée aux entreprises de moins de 20 salariés) ou encore un niveau de

ressources donnant accès à une bourse d'études ou à un crédit d'impôt. Comme le montrent ces exemples, l'hypothèse selon laquelle la variable à laquelle le seuil s'applique (par exemple l'âge, ou la taille de l'entreprise), communément appelée « variable de forçage » (ou *forcing variable*) n'aurait pas d'influence sur la variable de résultat de l'intervention, est en général peu crédible. Le départ en retraite va de pair avec une augmentation de l'âge qui a par lui-même de nombreuses conséquences sur l'état de santé, les habitudes de consommation, la vie sociale, etc. Les entreprises de grande taille interviennent dans des secteurs généralement distincts de ceux dans lesquels opèrent les PME, leur structure et leur activité sont souvent très différentes. Le niveau de revenu a évidemment un impact majeur sur de nombreuses décisions des ménages. Dans ces conditions, les deux groupes ainsi formés ne permettent pas de réaliser une évaluation de l'effet de l'intervention en comparant directement la valeur de la variable de résultat entre les bénéficiaires et les non-bénéficiaires.

L'application d'un seuil d'éligibilité produit en revanche une discontinuité soudaine dans la répartition des observations qui se situent à proximité du seuil : les observations dont la variable de forçage se trouve juste au-dessous du seuil bénéficient de l'intervention tandis que leurs voisins, dont la variable de forçage est située juste en-dessus, en sont exclus. La régression sur discontinuité exploite cette propriété en faisant l'hypothèse que les variations de faible ampleur de la variable de forçage autour du seuil résultent d'un pur aléa, similaire à un tirage au sort, qui détermine l'accès à l'intervention d'observations par ailleurs identiques. À proximité du seuil, l'affectation d'une observation au groupe de traitement s'apparente ainsi à une expérimentation randomisée. Lorsque les observations sont ordonnées en fonction de la valeur croissante de la variable de forçage, toute rupture dans la valeur de la variable de résultat une fois le seuil franchi peut, sous cette hypothèse, être interprétée comme une mesure de l'effet de l'intervention.

Dans sa forme la plus simple, la méthode de régression sur discontinuité mesure donc l'effet d'une politique en comparant la valeur moyenne de la variable de résultat dans le groupe des personnes bénéficiaires, par exemple celles dont le revenu ou l'âge est juste inférieur au seuil d'éligibilité, et la valeur moyenne de cette variable dans le groupe de contrôle comparable, formé des personnes dont le revenu ou l'âge est juste supérieur à ce seuil. L'hypothèse sous-jacente est que, pour des personnes ayant par ailleurs les mêmes caractéristiques du point de vue de la qualification, du niveau d'éducation ou du genre, celles situées juste en-dessous et au-dessus du seuil sont identiques. Cette mise en œuvre de la méthode nécessite donc de définir l'intervalle au sein duquel les observations sont conservées pour l'analyse. Cette « fenêtre d'observation » est choisie en réalisant un arbitrage entre la qualité de l'analyse statistique permise par un échantillon de plus grande taille et la fragilisation de cette hypothèse qui résulte d'un élargissement de l'intervalle. Imbens et Kalyanaraman (2012) proposent une méthode pour choisir de manière optimale l'amplitude de cet intervalle (bandwidth).

Cette régression sur discontinuité est dite stricte (« *sharp regression discontinuity design* ») lorsque l'affectation au groupe pouvant bénéficier de l'intervention ou du dispositif public est obligatoire et strictement déclenchée par la valeur de la variable de forçage. Si l'éligibilité est fondée, par exemple, sur un critère d'âge, et appliquée par une autorité qui dispose d'un recensement exhaustif de la population, alors la probabilité de bénéficier de l'intervention est égale à 1 dès lors que la condition d'âge est remplie; et cette probabilité est égale à 0 sinon, de sorte que l'affectation en fonction du seuil est un événement certain. Prenons l'exemple d'un programme de formation pour les personnes en recherche d'emploi ayant 25 ans révolus. Le principe est alors de comparer la moyenne de la variable de résultat (la variable dépendante, par exemple, le salaire d'embauche au moment du retour à l'emploi) pour les personnes en recherche d'emploi qui sont juste au-dessus du seuil d'âge, par

exemple âgées de 25 ou 26 ans, et le salaire moyen d'embauche pour celles qui sont âgées de 23 ou 24 ans, qui n'ont pu bénéficier du programme de formation.

La régression sur discontinuité floue (*fuzzy regression discontinuity design*) correspond à l'inverse aux situations dans lesquelles ce seuil est moins contraignant, si bien qu'il existe de part et d'autre du seuil des observations qui sont ou non bénéficiaires de l'intervention. Dans l'exemple du programme de formation destiné aux personnes en recherche d'emploi de 25 ans et plus développé ci-dessus, supposons que, dans une localité donnée, cette formation ne puisse être dispensée qu'à 100 personnes âgées de 25 ou 26 ans en raison de contraintes budgétaires, et que cette formation ne soit pas obligatoire, de sorte que seuls 80 de ces 100 personnes éligibles (soit 80% d'entre elles) acceptent effectivement de participer à la formation. L'agence locale pour l'emploi propose alors les 20 places restantes à des personnes de 23 ou 24 ans; qui sont également au nombre de 100. Seules 10 d'entre elles (soit 10%) acceptent de participer à la formation. Plutôt qu'à un changement soudain du statut vis-à-vis de l'intervention, la notion de discontinuité fait donc référence ici au « saut » que subit la probabilité de bénéficier de l'intervention lorsque le seuil d'éligibilité (25 ans) est franchi. L'objectif de l'approche est donc de mesurer l'effet de l'intervention en se restreignant à la variation de la variable de résultat qui résulte de ce « saut » dans la probabilité de bénéficier de l'intervention évaluée. Cette approche repose sur une hypothèse forte, appelée hypothèse de monotonicité (« *monotonicity* ») : cette hypothèse implique que parmi les chômeuses et chômeurs qui ne participent pas au programme de formation parce que leur âge est inférieur à 25 ans, il existe un sous-groupe d'individus qui accepteraient d'y participer si leur âge était de 25 ans révolus. En termes techniques, ces observations sont appelées « *compliers* », terme que l'on peut traduire par « personnes observantes », « qui obtempèrent » ou « qui se conforment à la prescription ». Par construction, la régression sur discontinuité floue mesure l'effet de l'intervention uniquement pour ce sous-groupe. Outre que ce sous-groupe peut être parfois de taille très restreinte, il exclut

deux groupes importants : les individus qui sont prêts à participer à l'intervention, quelle que soit la valeur de la variable de forçage (les « *always takers* »), ainsi que ceux qui ne souhaitent pas y participer en toutes circonstances (les « *never takers* »).

III. Deux exemples d'utilisation de cette méthode dans l'éducatif

Les variations de prix de l'immobilier entre quartiers reflètent la disposition à payer des contribuables pour l'ensemble des services et aménités (caractéristiques de l'environnement) auquel un logement donne accès. L'une de ces aménités est bien sûr la qualité de l'école de secteur à laquelle les enfants des résidents pourront accéder. Les tentatives d'estimation des effets de la qualité des écoles sur le prix des logements sont souvent peu convaincantes car les meilleures écoles ont tendance à se trouver dans les meilleurs quartiers. Ces deux aspects contribuent conjointement à une élévation du prix de l'immobilier et les évaluations qui ne tiennent pas suffisamment compte des caractéristiques du quartier tendent donc à surestimer la valeur des écoles qui y sont situées. Pour contourner cette difficulté, Black (1999) recourt à une application particulièrement originale de la méthode de la régression sur discontinuité stricte, en s'appuyant sur un seuil correspondant aux contours de la carte scolaire à Boston. L'étude consiste à estimer la valeur que les parents accordent à la qualité de l'école publique de secteur, en comparant les logements qui sont situés de part et d'autre des limites géographiques d'un secteur scolaire incluant les écoles publiques auxquelles les enfants sont affectés. Le fait que les résultats moyens obtenus par les élèves d'écoles de secteurs différents mais voisins varient parfois fortement, alors que les caractéristiques des logements situés des deux côtés des divisions scolaires changent par définition assez peu, permet d'isoler par discontinuité la relation entre les résultats scolaires (interprétés comme la qualité des écoles) et les prix

des logements. Les estimations suggèrent qu'une augmentation d'un point de la moyenne obtenue aux tests scolaires entraîne une augmentation de 1,3% à 1,6% du niveau des prix des logements situés à la limite du secteur scolaire.

L'étude de Matsudaira (2008) constitue un exemple de mise en œuvre de la méthode de la régression sur discontinuité floue, également appliquée à la réussite scolaire. L'étude exploite un ensemble de données administratives provenant d'un grand district scolaire des États-Unis. Dans ce district, les élèves accèdent à la classe suivante si leurs notes sont supérieures à des seuils prédéfinis. Les élèves obtenant des notes inférieures à ces seuils doivent participer à une école d'été de quatre à six semaines afin d'éviter le redoublement. Étant donné que les caractéristiques observées des élèves au voisinage des seuils sont presque identiques, les différences de résultats scolaires entre les élèves qui se situent juste en dessous et juste au-dessus des seuils peuvent être attribuées à l'impact causal de l'école d'été. L'échantillon est restreint aux élèves scolarisés entre le cours élémentaire de deuxième année (CE2, à l'âge de huit ans environ) et la classe de cinquième (à l'âge de 12 ans environ). Les résultats des élèves ont été enregistrés aux examens de mathématiques et de lecture aux printemps 2001 et 2002, ce qui donne un échantillon d'analyse de 338 608 élèves. Cependant, la régression sur discontinuité est ici floue : la relation entre les résultats aux tests de fin d'année et la fréquentation de l'école d'été n'est pas absolument obligatoire. Certains élèves dont les notes étaient inférieures aux seuils n'ont pas suivi cette école d'été, alors que d'autres, dont les notes étaient supérieures aux seuils, s'y sont inscrits. Seuls 38% des élèves de CE2 et de cours moyen de deuxième année (CM2) dont les notes en mathématiques étaient inférieures aux prérequis à la fin de l'année scolaire 2000-2001 se sont inscrits à l'école d'été de 2001. Les estimations provenant de la mise en œuvre de la régression sur discontinuité floue suggèrent que les notes des élèves « observants » de CE2 ont augmenté de 12,8% l'année suivante, alors que celles des élèves « observants » de CM2 participant à

cette école d'été ont augmenté de 24,1%. Les effets se sont avérés faibles pour les élèves « observants » de sixième, voire inexistants pour les élèves « observants » de cours moyen de première année (CM1) et de cinquième.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Pour que la discontinuité s'apparente à une expérimentation locale, il est important que la variable de forçage soit une donnée objective qui échappe au contrôle des populations concernées par l'intervention. S'il est possible de « manipuler » le positionnement de cette variable par rapport au seuil, l'affectation au groupe de traitement devient alors une variable de choix. L'exemple classique est celui d'une politique publique qui accorde des aides à l'emploi aux entreprises de moins de 20 salariés. La réaction naturelle de certaines entreprises dont l'effectif s'approche du seuil est d'intensifier le recrutement d'intérimaires, afin d'augmenter leur effectif sans que cette augmentation n'apparaisse dans les déclarations fiscales auxquelles elles sont soumises, de manière à continuer à bénéficier des aides à l'emploi. Pour détecter de telles manipulations du seuil, McCrary (2008) propose un test statistique simple, qui repose sur un raisonnement agrégé. Les entreprises qui emploient en réalité plus de 20 salariés (21 ou 22 salariés par exemple), mais dont la taille déclarée est inférieure à 20 salariés (soit 19 ou 20), vont faire croître artificiellement la proportion d'entreprises de moins de 20 salariés dans le secteur et simultanément faire diminuer la proportion d'entreprises employant 21 ou 22 salariés. L'existence de manipulations en réaction au seuil d'éligibilité a donc des conséquences sur la distribution de la taille des entreprises, qui peut être inspectée à l'aide d'un histogramme. En théorie, cet histogramme ne devrait pas faire apparaître de discontinuité juste avant et juste après le seuil de 20 salariés. Si tel était toutefois le cas, et cela peut être testé statistiquement, on pourrait alors suspecter un comportement de « manipulation » de la part des entreprises.

Pour éviter de réduire trop fortement la fenêtre d'observation utilisée, il est fréquent d'ajouter des variables explicatives autres que la variable de forçage : cela permet de tenir compte des variations de la variable de résultat qui sont dues à certaines caractéristiques observables. Le revenu, par exemple, a tendance à croître avec l'âge, de sorte qu'un élargissement de la fenêtre autour du seuil d'âge conduit à inclure des observations dont la variable de résultat change en raison des variations de revenu. La prise en compte de cet effet du revenu permet d'éliminer cette différence entre les groupes. Il est important de vérifier que la distribution de ces variables ne présente pas de discontinuité au voisinage du seuil considéré. Dans le cas contraire, cela signifie que l'intervention que l'on souhaite évaluer a des effets non seulement sur la variable de résultat mais aussi sur ces variables de contrôle. Prendre en compte ces variables dans l'analyse statistique conduit à des estimations biaisées de l'effet de l'intervention sur la variable de résultat, car ces variables sont elles-mêmes expliquées par la mise en œuvre de l'intervention.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

La principale difficulté à laquelle se heurtent la plupart des méthodes quasi-expérimentales est qu'elles reposent sur des hypothèses fortes, souvent remises en cause, quant à la comparabilité du groupe contrôle et du groupe cible après intervention. C'est ce qui conduit, par exemple, lorsque l'on souhaite mettre en œuvre la méthode des doubles différences, à s'assurer à la fois que la variable de résultat a connu par le passé la même évolution dans les deux groupes et que leurs caractéristiques observables sont semblables. La difficulté est la même lorsque l'on souhaite recourir à une méthode d'appariement d'échantillon (*matching*) : il convient pour ce faire de trouver des observations servant de groupe de contrôle qui présentent certes des caractéristiques observables similaires à celles des observations du groupe cible, mais

qui ont surtout une probabilité non nulle d'être éligibles à la politique évaluée. La méthode de régression sur discontinuité échappe à cette difficulté car elle repose sur un principe d'affectation aléatoire au sein de la sous-population qui se situe au voisinage du seuil. À l'instar d'une expérimentation contrôlée, la comparabilité entre les observations appartenant à chacun des deux groupes repose sur un argument statistique : si la taille de l'échantillon est suffisante, la distribution de toutes les caractéristiques qui sont pertinentes pour expliquer les variations de la variable de résultat est similaire entre les deux groupes.

Cette assimilation de la discontinuité à une expérimentation aléatoire est d'autant plus convaincante que l'intervalle à l'intérieur duquel elle est supposée se dérouler est étroit, ce qui conduit à restreindre l'effet mesuré à une sous-population très particulière, qui se caractérise par la proximité de la variable de forçage par rapport au seuil. La mesure fournie par cette expérimentation aléatoire locale est donc spécifique à cette sous-population. Dès lors que l'effet de l'intervention varie fortement entre différents sous-groupes, la mesure qui en découle est donc elle-même locale et seulement valable au voisinage du seuil retenu (ce qui correspond à un effet local moyen, ou LATE « *local average treatment effect* »). L'extrapolation des résultats obtenus à des sous-populations éloignées du seuil (qui définirait la « validité externe » de la mesure obtenue) n'a que très peu de pertinence. Cette limite de la méthode est encore amplifiée dans le cas d'une régression sur discontinuité floue, dont l'effet local est spécifique aux seules entités « observantes ». Ce manque de validité externe est d'autant plus problématique que les seuils sont souvent fixés en fonction du bénéfice attendu de l'intervention dans les populations ciblées. Un programme de formation à destination des personnes qui connaissent une situation de chômage de longue durée vise par exemple à contrecarrer les effets des pertes en capital humain dues à des épisodes de chômage prolongés. La fixation d'un seuil permettant de distinguer les épisodes de chômage de longue durée est en partie fondée sur le fait que cette perte de capital humain reste minime lorsque les épisodes sont suffisamment courts : l'estimation de l'effet d'un

tel programme par la méthode de la régression sur discontinuité revient donc à centrer l'analyse sur la partie de la population pour laquelle l'effet du programme est très probablement le moins fort, à savoir les personnes en recherche d'emploi dont les épisodes de chômage sont relativement plus courts.

Le lectorat intéressé trouvera d'excellentes synthèses sur la méthode de la régression sur discontinuité, par exemple, dans l'article de Lee et Lemieux (2010), ou bien dans l'ouvrage de Cattaneo, Idrobo et Titiunik (2019).

Quelques références bibliographiques pour aller plus loin

Black, Sandra E. 1999. « Do Better Schools Matter? Parental Valuation of Elementary Education ». *Quarterly Journal of Economics* 114 (2): 577-99. <https://doi.org/10.1162/003355399556070>

Cattaneo, Matias D.. et Idrobo, Nicolás. et Titiunik, Rocío. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations. Elements in Quantitative and Computational Methods for the Social Sciences*. Cambridge University Press. <https://doi.org/10.1017/9781108684606>

Imbens, Guido. et Kalyanaraman, Karthik. 2012. « Optimal Bandwidth Choice for the Regression Discontinuity Estimator ». *Review of Economic Studies*, 79 (3): 933-59. <https://doi.org/10.1093/restud/rdr043>

Lee, David S.. et Lemieux, Thomas. 2010. « Regression Discontinuity Designs in Economics ». *Journal of Economic Literature* 48 (2): 281-355. <https://doi.org/10.1257/jel.48.2.281>

Matsudaira, Jordan D. 2008. « Mandatory Summer School and Student Achievement ». *Journal of Econometrics*, The regression discontinuity design: Theory and applications, 142 (2): 829-50. <https://doi.org/10.1016/j.jeconom.2007.05.015>

McCrary, Justin. 2008. « Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test ». *Journal of Econometrics*, The regression discontinuity design: Theory and applications, 142 (2): 698-714. <https://doi.org/10.1016/j.jeconom.2007.05.005>

Des ressources pour mettre en œuvre cette méthode avec les logiciels Stata et R

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press: New Haven and London. Disponible en libre accès sur le site <https://mixtape.scunning.com/index.html>

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*, Chapitre 20. Chapman and Hall/CRC Press: Boca Raton, Florida. Disponible en libre accès sur le site <https://theeffectbook.net/ch-RegressionDiscontinuity.html>

4. Méthodes d'appariement

PAULINE GIVORD

Résumé

L'appariement est une méthode quantitative d'évaluation *ex post* dans laquelle, en l'absence d'expérimentation directe, on reconstitue une situation contrefactuelle en comparant les situations de bénéficiaires d'une intervention à celles de non bénéficiaires aux caractéristiques très proches. Cette méthode est notamment utile pour évaluer l'impact d'un programme sur l'ensemble d'une population, lorsqu'il existe des données suffisamment précises pour comparer les personnes bénéficiaires et non bénéficiaires.

Mots-clés : Méthodes quantitatives, évaluation *ex post*, effet propre, score de propension, support commun

I. En quoi consiste cette méthode?

Les méthodes d'appariement font partie des principales méthodes d'évaluation quantitative *ex post*, visant à mesurer l'effet d'un dispositif ou d'un programme (par exemple un programme de formation pour les demandeurs et demandeuses d'emploi, ou des aides localisées sur certains territoire) sur la situation des bénéficiaires. Comme la plupart des méthodes d'évaluation quantitatives, l'enjeu est d'estimer l'effet propre (*causal*) du dispositif sur la situation des bénéficiaires (par exemple, un retour à l'emploi après une formation, ou l'activité économique du territoire visé). L'objectif des méthodes d'appariement (*matching*) est d'estimer cet effet propre du dispositif en

comparant la situation des bénéficiaires du dispositif avec celle de personnes qui n'en ont pas bénéficié, mais qui ont des caractéristiques si proches qu'il aurait été possible qu'ils en bénéficient. L'observation de ces non bénéficiaires est supposée donner une idée de la situation « contrefactuelle », celle qu'aurait connue les bénéficiaires en l'absence du dispositif.

L'enjeu est ici de réduire les effets de sélection qui peuvent se produire quand on souhaite estimer l'effet d'un dispositif. En général, les bénéficiaires du dispositif n'ont pas été désigné·e·s par hasard, et ils et elles ont des caractéristiques spécifiques qui peuvent expliquer par elles-mêmes une évolution plus ou moins favorable, même en l'absence du dispositif évalué. Par exemple, l'évaluation d'un programme de formation à destination des personnes les plus éloignées de l'emploi ne peut se faire simplement en comparant les chances de retour à l'emploi des personnes formées avant ou après la formation, au risque de sous-estimer l'effet de la formation qui ciblerait des personnes ayant plus de difficultés. Il n'est pas non plus possible de comparer les taux de retour à l'emploi des personnes formées avec celui de l'ensemble des personnes non formées : ces dernières sont trop différentes pour que leur situation vis-à-vis de l'emploi soit un reflet probable de la situation qu'aurait connue les stagiaires en l'absence de formation.

Le principe des méthodes d'appariement est de restreindre la comparaison des bénéficiaires aux non bénéficiaires comparables. Précisément, on apparie chaque bénéficiaire du programme ou dispositif à évaluer avec un·e ou des non bénéficiaires « jumeaux » ou « jumelles », au sens où ils et elles ont des caractéristiques individuelles très proches dans toutes les dimensions pouvant influencer à la fois sur le fait de bénéficier du dispositif et sur la situation ultérieure. Dans l'exemple de l'estimation de l'impact du stage de formation sur les chances de retour à l'emploi, on compare pour chaque stagiaire par exemple le fait d'avoir retrouvé un emploi dans l'année suivant l'entrée en formation avec les chances de retour à l'emploi de personnes identiques ou tout au moins les plus proches à ce ou cette stagiaire à la date de l'entrée en formation

dans les dimensions considérées comme importantes pour le retour à l'emploi. L'effet moyen de la formation pour les stagiaires est obtenu en faisant la moyenne de l'ensemble de ces comparaisons menées pour chaque bénéficiaire.

En principe, on souhaite apparier sur le plus grand nombre de dimensions possibles, pour se prémunir du risque de manquer une caractéristique importante, et dont la non prise en compte dans les comparaisons conduirait à des estimations incorrectes de l'effet propre du dispositif. Cependant, plus les dimensions sur lesquelles on souhaite apparier sont nombreuses et plus il sera difficile de trouver pour chaque bénéficiaire des non bénéficiaires exactement identiques dans toutes ces dimensions. Dans l'exemple de l'évaluation d'un programme de formation, il pourra ainsi être pertinent d'apparier sur l'âge, le niveau de diplôme, l'ancienneté au chômage et l'expérience passée (par exemple le nombre d'épisodes de chômage antérieur), l'expérience professionnelle passée (par exemple la qualification de l'emploi), le type d'emploi recherché, la mobilité possible, qui sont autant de variables qui peuvent influencer tant sur le choix de se former que sur le retour à l'emploi (indépendamment de cette formation). Faire un appariement exact sur chacune de ces dimensions signifie qu'on doit trouver pour chaque stagiaire de la formation professionnelle une personne ayant des caractéristiques exactement identiques dans l'ensemble de ces dimensions : plus le nombre de variables est élevé, et moins il est probable, surtout si le nombre d'observations est faible, de trouver un « jumeau » parfait ou une « jumelle » parfaite.

Une réponse fréquemment utilisée est d'apparier non pas sur l'ensemble de ces caractéristiques, mais sur un résumé de celles-ci fourni par le « score de propension ». Celui-ci correspond à la probabilité d'être bénéficiaire, conditionnellement aux dimensions retenues comme importantes pour l'appariement. Cela signifie que l'estimation se fait en deux étapes. Dans un premier temps, on estime ce score de propension, c'est-à-dire comment les différentes dimensions prédisent l'entrée en formation, ce qui permet pour chaque observation de définir une probabilité *a priori* d'être bénéficiaire, en fonction de ses caractéristiques.

Dans notre exemple, on estimera la probabilité d'entrer en formation comme une fonction de l'âge, du diplôme, etc. On utilisera cette estimation pour calculer pour chaque personne, stagiaire de la formation ou non, sa « propension » à entrer en formation, c'est-à-dire la probabilité prédite en fonction de ces caractéristiques individuelles. Les valeurs du score de propension se situent en général strictement entre zéro et un (sauf condition d'exclusion particulière, il est rare qu'une personne n'ait aucune chance d'entrée en formation, et inversement il est peu vraisemblable qu'une des caractéristiques utilisées se traduise automatiquement par une entrée en formation). Leurs répartitions se recouvrent entre les bénéficiaires et les non bénéficiaires. Si les personnes qui ont *a priori* une probabilité élevée d'entrer en formation sont plus nombreuses parmi les personnes effectivement entrées en formation, certaines ne le font pas et peuvent servir de comparaison. Inversement, certaines personnes ayant *a priori* une propension faible à entrer en formation peuvent néanmoins choisir de se former – et il sera aussi possible de les comparer avec des personnes qui ne se sont pas formées, ayant également une faible propension à le faire. On peut montrer que lorsque l'on utilise un appariement sur le score de propension, les caractéristiques importantes sont en moyenne identiques entre le groupe des bénéficiaires et les non bénéficiaires.

Que l'appariement se fasse sur une seule dimension (le score de propension), ou sur plusieurs d'entre elles, il est difficile d'avoir des valeurs exactement identiques pour l'appariement : celui-ci se fait donc en utilisant les « plus proches voisin-e-s » des bénéficiaires, c'est-à-dire les non bénéficiaires qui se rapprochent le plus du ou de la bénéficiaire selon les dimensions retenues (ou selon le score de propension). Il existe ensuite plusieurs variantes notamment sur le nombre de voisin-e-s qu'on retient (on peut préférer en retenir plusieurs pour éviter de comparer par malchance avec un-e non bénéficiaire dont le comportement serait atypique) et la distance maximale qu'on autorise entre le ou la bénéficiaire et les comparaisons (des voisin-e-s trop éloigné-e-s étant par définition moins adapté-e-s pour la comparaison).

Quelle que soit la méthode d'appariement utilisée, il est nécessaire pour l'appliquer de disposer de données individuelles permettant de décrire finement la situation et les caractéristiques individuelles, et de très nombreuses observations pour avoir plus d'assurance de pouvoir trouver des voisin·e·s proches.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Les méthodes d'appariement permettent d'estimer *ex post* l'effet d'un programme sur les bénéficiaires, sur un ensemble de dimensions objectivables et mesurables. Elles permettent par exemple de répondre à des questions du type : des demandeur·euse·s d'emploi qui ont choisi de se former (au risque d'interrompre une recherche d'emploi) ont-ils ou elles une probabilité de retour à l'emploi durable *in fine* plus élevée que des demandeur·euse·s d'emploi qui ne se forment pas? Cette formation leur permet-elle d'espérer un niveau de rémunération plus élevée? Quels demandeur·euse·s d'emploi bénéficient le plus de la formation?

Il s'agit donc de mesurer des écarts entre la situation qui a été effectivement vécue par les bénéficiaires d'un programme, et une situation « contrefactuelle », qui aurait prévalu en l'absence de ce programme. En général, ces méthodes sont adaptées pour évaluer l'impact de la mise en œuvre d'un programme (par rapport à une situation où ce programme n'existerait pas), mais le sont moins pour mesurer l'effet des différentes modalités de cette mise en œuvre (dans notre exemple, plusieurs dispositifs plus ou moins intensifs de formation des demandeurs d'emploi).

III. Deux exemples d'application : politiques actives de l'emploi et exonérations fiscales territoriales

Les méthodes d'appariement sont très classiquement utilisées pour évaluer les effets des mesures dites « actives » de l'emploi (formation, aides à la recherche d'emploi, etc.), notamment depuis l'étude méthodologique d'Heckman, Ichimura et Todd (1997). Cette méthode a été utilisée par exemple pour étudier une politique d'emploi active en Suède (Sianesi, 2004), des programmes de formation en Allemagne ou plus récemment les formations à destination des personnes en recherche d'emploi en France (Chabaud *et al.*, 2022).

Un autre exemple est l'évaluation des effets des Zones franches urbaines (ZFU), dispositifs d'exonérations fiscales et sociales prévues pour favoriser l'implantation des entreprises dans des zones urbaines défavorisées, à l'image des *Enterprises Zones* mises en place aux États-Unis à partir des années 1980. Givord, Rathelot et Sillard (2013) s'intéressent aux effets de ces exonérations sur l'installation des entreprises et l'évolution de l'emploi dans les quartiers ciblés, en comparaison avec d'autres quartiers initialement très proches (voir aussi Malgouyres et Py, 2016). Ces études suggèrent un effet positif des zones sur l'emploi et l'activité économique, mais au détriment des zones immédiatement voisines. Une autre étude suggérait par ailleurs des effets non persistants au-delà de la durée des exemptions (Givord *et al.*, 2022).

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

La validité des méthodes d'appariement dépend de manière cruciale de la manière dont elles peuvent corriger des effets de sélection, et donc des informations disponibles pour comparer les bénéficiaires et les non-

bénéficiaires. Il faut avoir une assurance que le processus de sélection dans le dispositif ne se fasse pas en fonction de variables qui ne sont pas disponibles dans les données (par exemple, les résultats d'un entretien de motivation utilisé pour entrer dans un programme de formation, qui viserait à mesurer des dimensions peu objectivables et donc non disponibles pour un regard extérieur). Le fait de disposer d'informations individuelles sur la variable d'intérêt sur le passé (par exemple, la trajectoire professionnelle antérieure à l'entrée en formation) est en général considéré comme indispensable pour éviter de capter des effets de sélection : les méthodes d'appariement sont dans ce cas combinées avec des « différences de différences ». Ensuite, il est nécessaire qu'il soit possible d'apparier l'ensemble des bénéficiaires avec des non-bénéficiaires (on parle de « support commun »). Cette dernière condition signifie notamment qu'il existe un certain aléa dans le fait de bénéficier du programme : si celui est totalement déterministe en fonction des caractéristiques observables (par exemple un programme systématiquement proposé aux jeunes non diplômé·e·s, qui excluraient en revanche les personnes au-delà d'un seuil d'âge ou de revenu), il ne sera pas possible d'apparier les bénéficiaires.

Enfin, les méthodes d'appariement fournissent une estimation statistique, et donc comme telles ne permettent pas de mesurer avec une totale certitude la « vraie » valeur de l'effet mais seulement une approximation dont on peut quantifier la précision, c'est-à-dire le degré de confiance avec laquelle on peut utiliser cette estimation. Cette précision se mesure via l'écart-type (plus celui-ci est faible, et plus il est possible d'avoir confiance dans le fait que le « vrai » effet est proche de la valeur estimée) ou encore un intervalle de confiance, qui correspond à l'intervalle de valeurs au sein duquel le vrai effet se trouve avec une probabilité déterminée : par exemple, l'intervalle de valeur où la vraie valeur de l'effet se trouve avec une probabilité de 95% (plus l'intervalle de confiance est réduit, et plus la valeur estimée est connue avec précision). Cette mesure de précision est notamment utilisée pour vérifier que l'effet du dispositif

évalué est « significatif » ou « significativement différent de zéro », c'est-à-dire qu'on peut dire avec une certaine assurance que le programme a effectivement un effet strictement positif, ou strictement négatif.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

L'un des atouts des méthodes d'appariement est de pouvoir estimer les effets en « population générale », c'est-à-dire sur l'ensemble de la population (à condition de disposer de suffisamment d'observations pour pouvoir trouver des comparaisons et que l'assignation au dispositif comporte suffisamment d'aléas pour que l'on puisse disposer de bénéficiaires sur l'ensemble). Cela peut constituer un atout par rapport à la plupart des méthodes d'évaluation quantitative *ex post* ne permettent d'estimer sans biais un effet causal que sur des populations « marginales » : par exemple, les personnes autour d'un seuil d'éligibilité pour les régressions sur discontinuités ou les personnes qui sont sensibles au signal donné par un instrument.

En revanche, les méthodes d'appariement peuvent ne pas être suffisantes pour corriger des biais de sélection. Les estimations sont très sensibles au choix des variables utilisées pour l'appariement, et il est en général difficile de faire confiance à des estimateurs en l'absence de mesures individuelles passées de la variables d'intérêt.

Quelques références bibliographiques pour aller plus loin

- Biewen, Martin. et Fitzenberger, Bernd. et Osikominu, Aderonke. et Paul, Marie. 2014. « The Effectiveness of Public-Sponsored Training Revisited: The Importance of Data and Methodological Choices ». *Journal of Labor Economics*, 32: 837-897.
- Fitzenberger, Bernd. et Völter, Robert. 2007. « Long-run effects of training programs for the unemployed in East Germany ». *Labour Economics*, 14 (4): 730-755.
- Givord, Pauline. et Rathelot, Roland. et Sillard, Patrick. 2013. « Place-based tax exemptions and displacement effects: An evaluation of the Zones Franches Urbaines program ». *Regional Science and Urban Economics*, 43(1): 151-163.
- Givord, Pauline. et Quantin, Simon. et Trevien, Corentin. 2018. « A long-term evaluation of the first generation of French urban enterprise zones ». *Journal of Urban Economics*, n° 105(C): 149-161.
- Heckman, James. et Ichimura, Hidehiko. et Todd, Petra. 1997. « Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme ». *Review of Economic Studies*, 64(4): 605-654.
- Lechner, Martin. 2002. « Program Heterogeneity And Propensity Score Matching: An Application To The Evaluation of Active Labor Market Policies ». *The Review of Economics and Statistics*, vol. 84, n°2: 205-220.
- Malgouyres, Clément. et Py, Loriane. 2016. « Les dispositifs d'exonérations géographiquement ciblées bénéficient-ils aux résidents de ces zones? État des lieux de la littérature américaine et française ». *Revue économique*, 67: 581-614.

Sianesi, Barbara. 2004. « An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s ». *Review of Economics and Statistics*, 86: 133-155.

5. Microsimulation

MATHIAS ANDRÉ

Résumé

La microsimulation est une méthode quantitative permettant d'estimer l'impact attendu d'une intervention (par exemple, la modification d'un taux d'imposition) et d'en décrire les effets (gagnant·e·s, perdant·e·s, enveloppe budgétaire, effet sur les indicateurs d'inégalités). Elle repose sur une prise en compte des caractéristiques de la population ciblée (par exemple en termes d'âge, de revenu, etc.) et modélise les effets des politiques publiques concernant cette population. En raison de la diversité des situations qu'elle permet d'intégrer, cette technique fournit des résultats plus précis et plus complets que les estimations à partir d'un raisonnement moyen ou agrégé de type individu représentatif. Son développement a été favorisé par l'amélioration de la puissance des calculs informatiques, la multiplication des informations statistiques (enquêtes ou données administratives). C'est un outil essentiel de l'évaluation *ex ante* de l'impact des politiques publiques qui peut également être utilisé pour l'évaluation *ex post*.

Mots-clés : Méthodes quantitatives, modélisation, microsimulation statique/dynamique, démographie, politiques socio-fiscales, retraites

I. En quoi consiste cette méthode?

Les méthodes de microsimulation sont des outils développés par des unités de recherche ou des administrations dans un objectif de modélisation des agent·e·s économiques, principalement des individus ou

des entreprises, à des fins d'évaluation des politiques publiques. Elles ont été développées pour dépasser les limites de l'analyse macroéconomique avec un seul agent représentatif de l'économie (Orcutt, 1957). Le principe général repose sur la représentation de l'économie sous la forme d'un ensemble d'unités élémentaires (les individus, par exemple) caractérisées par un certain nombre de variables (l'âge, le statut marital, la taille de la famille, le revenu, par exemple). Par distinction avec une modélisation qui prendrait appui sur un individu moyen, cette description permet de mesurer la disparité des situations et d'en modéliser les évolutions. Par exemple, les méthodes dynamiques qui cherchent à modéliser la démographie vont faire vieillir, mourir et naître des individus, modélisation qui pourra être utile pour évaluer les systèmes de retraite. D'autres modèles peuvent étudier l'effet des changements des impôts ou des prestations sociales sur le revenu disponible des ménages. C'est l'objet de la microsimulation socio-fiscale. Autrement dit, le microsimulateur va s'appuyer sur une variété de données observées et simuler des changements de situations, unité par unité, soit par l'intermédiaire de relations déterministes (si les prestations familiales augmentent, les ménages avec enfant voient leurs revenus augmenter) ou probabiliste (chaque année, des enfants naissent avec une certaine probabilité). La microsimulation est dite dynamique si elle intègre des phénomènes tels que les évolutions démographiques (vieillesse, fécondité, mortalité) ou les ajustements sur différents marchés (l'emploi, le commerce, etc.); elle est dite statique autrement.

La microsimulation s'appuie sur les outils informatiques pour assurer la modélisation de la variété socio-économique. Elle fait appel à des logiciels statistiques dédiés (R ou Python par exemple) pour traiter à la fois les données individuelles et pour l'écriture du *modèle* à proprement parler. Celui-ci *simule* les situations à partir des variables observées et s'appuie notamment sur une programmation à partir de la législation qui applique les barèmes en vigueur aux situations des individus (comme l'âge de départ à la retraite ou le calcul de l'impôt sur le revenu). La législation existante est ainsi modélisée par défaut et est comparée aux réformes

étudiées, également modélisées pour l'exercice d'évaluation. Concrètement, un modèle de microsimulation va s'appuyer sur trois briques : le sujet traité, les données utilisées et un « calculateur », c'est-à-dire le cœur du code décrivant les évolutions ou les effets des phénomènes socio-économiques étudiés. Dans le cadre le plus répandu de la microsimulation socio-fiscale statique, cela permet par exemple de simuler les effets directs d'une réforme sous la forme d'effets totaux agrégés (le budget d'une modification fiscale par exemple), les effets directs sur les ménages (le nombre de gagnants ou perdants ainsi que les gains et pertes moyennes) et les effets redistributifs (mesurés par les changements dans les indicateurs d'inégalités ou la description des populations concernées). Certains travaux plus élaborés visent à tenir compte des comportements des agent·e·s en réaction aux politiques simulées.

Le principe initial des méthodes de microsimulation date des années 1960 (Orcutt, 1960) mais leur développement dans les années 1980 s'est surtout appuyé sur la généralisation des bases de données d'enquêtes représentatives des populations de la part des instituts statistiques et d'un plus grand usage des données administratives dans les méthodes d'évaluation quantitative dans les années 1990. Avec l'amélioration de la puissance des calculs informatiques et l'accès à une grande variabilité d'informations individuelles, les travaux universitaires ou administratifs se sont généralisés depuis les années 2000. Au Royaume-Uni et aux États-Unis et dans une moindre mesure en France, ce sont des outils installés dans le débat public, notamment dans le cadre de l'évaluation des systèmes de retraites ou des propositions socio-fiscales, lors des débats budgétaires par exemple.

Méthodes installées et reconnues, les techniques de microsimulation traitent une grande variété de sujets : la fiscalité et les transferts socio-fiscaux, les systèmes de retraites, les dépenses de santé et le système d'assurance-maladie, les politiques environnementales ou énergétiques, l'emploi et les trajectoires professionnelles, les choix d'éducation, la démographie, la dépendance, etc.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

La microsimulation est très régulièrement mobilisée pour l'évaluation *ex ante* des réformes socio-fiscales, éducatives ou environnementales. Ses résultats alimentent les études d'impact des lois ou des travaux d'études publiés par les équipes de microsimulation. La microsimulation s'appuie sur le calcul, la « simulation », de situations fictives. Le cœur de l'évaluation permise par la microsimulation s'appuie sur la comparaison de situations contrefactuelles sous la forme « avec ou sans réforme ». La simulation d'une législation nouvelle ou d'évolutions modifiées par des changements socio-démographiques permet de comparer deux situations. Pour évaluer une modification fiscale par exemple, le modèle compare les situations individuelles en appliquant ou non cette réforme. Par différence, il est alors possible d'estimer les gains et les pertes et d'en déduire les totaux (coûts ou recettes) et les distributions associées. La microsimulation fournit la population concernée : par exemple, les plus aisé·e·s, les retraité·e·s, les familles monoparentales, etc. Les scénarios prospectifs peuvent être nombreux et fournir alors à la fois une aide à la décision au législateur et une évaluation *ex ante* des politiques publiques.

En France, les services ministériels s'appuient ainsi sur des modèles afin de construire les politiques gouvernementales. La Direction générale du Trésor utilise le modèle Saphir pour ce qui concerne les prestations sociales monétaires et les prélèvements directs comme l'impôt sur le revenu par exemple. La Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (Drees), service statistique du ministère de la santé et des affaires sociales, développe plusieurs modèles, comme Trajectoire sur les retraites, OMAR sur les dépenses de santé, Autonomix pour la dépendance ou INES (co-développé avec l'Insee et la Cnaf et en accès libre depuis 2016) sur les politiques socio-fiscales. Le modèle Prometheus du ministère de l'environnement étudie par exemple des dépenses de chauffage et de transport des ménages français.

C'est également une tradition ancienne des économistes du campus Paris-Jourdan avec le modèle Sysiff développé dans les années 1970-1980 au laboratoire Delta (un laboratoire ancêtre de *Paris School of Economics* – PSE), la contribution au modèle EUROMOD utilisé par Eurostat et différents laboratoires de recherche en Europe, le simulateur fiscal de Camille Landais, Thomas Piketty et Emmanuel Saez sur lequel repose l'ouvrage grand public Landais, Piketty, Saez (2011). Actuellement, l'Institut des politiques publiques développe les modèles TaxIPP (socio-fiscal) ou PensIPP (retraites). Au Royaume-Uni par exemple, le budget est évalué par un institut (*Institute for fiscal studies*) en amont des débats au Parlement sur la base de modèles de microsimulation. Aux États-Unis, le modèle TaxSim est développé par le *National Bureau of Economic Research* (NBER) et accessible aux chercheurs.

Néanmoins, des usages *ex post* de la microsimulation sont également possibles. Le principe est identique aux méthodes *ex ante* mais ils s'appliquent à des politiques publiques effectivement appliquées. L'avantage de cet usage est qu'il ne requiert plus d'effectuer des hypothèses sur l'état de l'économie; il s'agit alors d'appliquer les simulations sur des données observées sur la période d'étude et de comparer la situation contrefactuelle à la situation réelle. La microsimulation *ex post* fait notamment l'objet d'études dans le domaine socio-fiscal, que décrit la section suivante.

III. Deux exemples d'utilisation de cette méthode : politiques socio-fiscales et politiques de retraites

Le modèle Ines, à l'époque co-développé par l'Insee et la Drees, a par exemple participé activement à la création de la prime d'activité en 2016, ainsi que du Revenu de solidarité active (RSA) en 2009. Il s'est d'abord agi de construire cette prestation sur la base des objectifs du législateur. De nombreux scénarios ont été chiffrés. Une fois le principe de la prestation

et l'enveloppe budgétaire cible fixés, la microsimulation a permis de fixer le montant du barème, ici le bonus individuel, correspondant aux critères. C'est avec ces moyens qu'est ensuite rédigée l'étude d'impact de la loi. La microsimulation a ainsi permis de construire le barème d'une politique sociale de premier plan.

Deux cas d'utilisation répandue de la microsimulation pour l'évaluation des politiques publiques sont l'étude du système de retraites (voir, Cheloudko et Martin, 2020) et celle des réformes socio-fiscales (Fredon et Sicsic, 2020). Sur ce sujet, l'équipe du modèle Ines publie chaque année une évaluation des réformes socio-fiscales de l'année passée et en dresse le bilan redistributif sur la base d'une méthodologie définie précisément (André et al., 2015). L'étude la plus récente (Buresi et al., 2022) fait ainsi état que « les nouvelles mesures sociales et fiscales intervenues en 2020 et 2021, une fois pleinement montées en charge, augmentent de 1,1% le niveau de vie des personnes résidant en France métropolitaine par rapport à une situation sans leur mise en œuvre. Le gain moyen atteint 280 euros par an et par personne : 240 euros pour les mesures de 2020 et 40 euros pour celles de 2021. Cette hausse profite surtout à la moitié la plus aisée de la population, particulièrement concernée par les principales réformes pérennes mises en œuvre ».

Dans un exercice similaire, l'Institut des politiques publiques et l'OFCE publient des évaluations des réformes, *ex ante* dans le cadre du budget (Fabre et al., 2020) ou parfois de façon *ex post* sur un quinquennat (Madec, Plane et Sampognaro, 2022). Ces analyses sont souvent reprises dans le débat public, que ce soit dans le champ médiatique avec de nombreux articles de presse qui s'appuient dessus ou dans le cadre de l'activité parlementaire avec des citations dans les rapports ou dans les déclarations des représentant-e-s politiques. C'est également le cas de la réforme de l'imposition des hauts patrimoines avec la transformation de l'impôt de solidarité sur la fortune (ISF) en impôt sur la fortune immobilière (IFI) qui a vu un débat s'installer sur la population effectivement concernée par cette réforme et les montants en jeu.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

La qualité d'une méthode de microsimulation repose à la fois sur la qualité des données sous-jacentes et sur celle du modèle utilisé. La représentativité de l'enquête ou des bases administratives assure la validité externe des résultats, c'est-à-dire la capacité du modèle à estimer les effets sur l'ensemble de la population cible. La richesse des variables de l'enquête emploi produite par l'Insee permet par exemple des représentations selon différents angles (statut d'activité, diplôme, etc.) alors que la granularité fine des données administratives fournit des échantillons larges afin de représenter les résultats sur des populations précises (le 1% les plus aisé·e·s par exemple).

Une méthode systématique de comparaison des résultats des modèles avec des sources externes garantit la qualité des simulations. Le modèle Ines fait ainsi l'objet d'une note annuelle dite « de validation ». Chaque prestation et impôt est comparé aux agrégats administratifs réels. Pour l'impôt sur le revenu par exemple, le nombre d'individus imposables, le total et le montant moyen sont des critères de qualité des simulations. Une documentation précise ainsi que la disponibilité du code source au format ouvert, c'est-à-dire accessible à tous, sont aussi un gage de transparence et donc de qualité d'un modèle de microsimulation.

Enfin, des disparités peuvent apparaître entre les résultats de différents modèles. La confrontation des résultats, ainsi que l'explication des écarts, permettent de juger des avantages et inconvénients des différents modèles (André et Sicsic, 2020).

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Les principaux atouts de la microsimulation résident dans ce qui a motivé sa création : les modèles permettent de rendre compte de la grande diversité des situations individuelles. L'écriture de la législation de façon intégrée permet de simuler des « effets détaillés de politiques dont les règles dépendent d'un grand nombre de caractéristiques individuelles, très souvent non linéaires, par exemple en raison d'effets de seuil ou de plafond » (Blanchet, 2020), comme par exemple les allocations logement ou l'impôt sur le revenu.

Les principales limites reposent sur l'exercice sans effet d'équilibre : les unités des modèles sont supposées ne pas modifier leur comportement (surtout dans les modèles statiques) ou d'interagir autrement que par les hypothèses limitées du modèle (démographiques ou choix de retraites dans les modèles dynamiques). Les évaluations sont ainsi qualifiées « de premier tour », c'est-à-dire qu'elles ne prennent pas en considération les effets de bouclage macroéconomique (comme les effets sur le marché du travail par exemple) ou les réactions comportementales (comme les ajustements de l'épargne ou de la consommation par exemple). Néanmoins, la prise en compte des comportements de non recours à certaines prestations sociales est parfois prise en compte dans les évaluations et dans les modèles statiques et constitue ainsi une intégration du comportement des ménages face aux politiques socio-fiscales.

Certaines études visent à tenir compte de ces limites et intègrent des effets de comportement à la suite des estimations des modèles de microsimulation (Paquier et Sicsic, 2021) ou d'effets de second tour (André et Biotteau, 2021).

Références bibliographiques

- André, Mathias. et Biotteau, Anne-Lise. 2021. Effets de moyen terme d'une hausse de TVA sur le niveau de vie et les inégalités : une approche par microsimulation. *Économie et Statistique*, n°522-523.
- André, Mathias. et Cazenave, Marie-Cécile. et Fontaine, Maëlle. et Fourcot, Juliette. et Sireyjol, Antoine. 2015. Effet des nouvelles mesures sociales et fiscales sur le niveau de vie des ménages : méthodologie de chiffage avec le modèle de microsimulation Ines. *Insee, Documents de travail*, n°F1507.
- André, Mathias. et Sicsic, Michaël. 2020. Évaluation des effets redistributifs des réformes socio-fiscales : comment s'y retrouver?, blog de l'Insee. <https://blog.insee.fr/evaluation-des-effets-redistributifs-des-reformes-socio-fiscales-comment-sy-retrouver/>
- Blanchet, Didier. 2020, Des modèles de microsimulation dans un institut de statistique : Pourquoi, comment, jusqu'où?, *Courrier des statistiques*, n°4.
- Buresi, Gabriel. et Cornetet, Jules. et Cornuet, Flore. et Doan, Quynh-Chi. et Dufour, Camille. et Trémoulu, Raphaël. 2022. Les réformes sociofiscales de 2020 et 2021 augmentent le revenu disponible des ménages, en particulier pour la moitié la plus aisée. *France portrait social, Insee références*.
- Cheloudko, Pierre. et Martin, Henri. 2020, Une décennie de modélisation du système de retraite – La genèse du modèle de microsimulation TRAJECTOIRE, *Courrier des statistiques*, n°4.
- Fabre, Brice. et Guillouzouic, Arthur. et Lallemand, Chloé. et Leroy, Claire. 2020. Budget 2020 : quels effets pour les ménages?, Note IPP n°49.
- Fredon, Simon. et Sicsic, Michaël. 2020, Ines, le modèle qui simule l'impact des politiques sociales et fiscales, *Courrier des statistiques*, n°4.

Landais, Camille. et Piketty, Thomas. et Saez, Emmanuel. 2011. *Pour une révolution fiscale*, La Découverte.

Madec, Pierre. et Plane, Mathieu. et Sampognaro, Raul. 2022. Une analyse macro et microéconomique du pouvoir d'achat des ménages en France : Bilan du quinquennat mis en perspective. *OFCE Policy Brief*, 104: 1-18.

Paquier, Félix. et Sicsic, Michaël. 2021. Effets des réformes 2018 de la fiscalité du capital des ménages sur les inégalités de niveau de vie en France : une évaluation par microsimulation. *Économie et Statistique*, n°530-531.

Quelques références bibliographiques pour aller plus loin

Des numéros de revues dédiés à la microsimulation

Courrier des statistiques n°4, avril 2020;

Économie et statistiques (n°481-482, 2015) et *Revue économique* (vol. 67, 2016);

Économie et prévision (n°160-161, 2003).

Des références généralistes dressant un panorama de la méthode

Bessis, Franck. et Cotton, Paul. 2021. La réforme, le chiffrage, son modèle et ses données, *Politix* 2021/2 (n°134).

Bourguignon, François. et Landais, Camille. 2022. Micro-simuler l'impact des politiques publiques sur les ménages : pourquoi, comment et lesquelles?, *Les notes du conseil d'analyse économique*, n°74, septembre 2022.

Legendre, François. L'émergence et la consolidation des méthodes de microsimulation en France. *Économie et Statistique*, n°510-511-512 – 2019, 201-217.

O'Donoghue, Cathal (éd). 2014. *Handbook of Microsimulation Modelling*, Emerald Publishing Ltd.

6. Expérimentation en laboratoire

LOU SAFRA

Résumé

L'expérimentation en laboratoire permet de mesurer directement les attitudes et comportements des individus et d'évaluer l'effet causal d'une variable sur ces comportements et attitudes. Pour cela, les individus sont mis en situation de réaliser un certain nombre de tâches dont l'on contrôle le plus d'éléments possibles (tels que la durée de la tâche et le type d'informations données aux participant·e·s). Cette démarche peut aider à anticiper *ex ante* la façon dont des individus vont réagir à une intervention, ou encore être utilisée *ex post* pour mesurer les changements de comportement à la suite d'une intervention. Elle est notamment utile pour révéler des biais de comportement non conscients.

Mots-clés : Méthodes quantitatives, méthode intra-participant·e·s/inter-participant·e·s, expérimentation en laboratoire, effet causal, comportements, attitudes, biais de comportement non conscients, validité interne/externe, réponse automatique/non-automatique

I. En quoi consiste cette méthode?

De façon simple, lors d'une expérimentation en laboratoire, des participantes et participants réalisent une tâche donnée, créée dans le but de mesurer leur comportement. La première étape de l'expérimentation en laboratoire est donc d'établir un protocole expérimental permettant de mesurer le comportement de l'individu. Classiquement, ces expérimentations impliquent la réalisation d'une tâche sur ordinateur, qui permettra de mesurer non seulement les choix des participant·e·s mais également d'autres données pouvant se révéler particulièrement informatives comme le temps de réponse. Ces tâches peuvent aussi bien viser à mesurer les préférences des participant·e·s et leurs perceptions que la façon dont ils et elles apprennent ou raisonnent. Ainsi, les expérimentations en laboratoire sont particulièrement utilisées par les domaines qui s'intéressent directement aux comportements et aux perceptions des individus comme les sciences cognitives, la psychologie, notamment la psychologie sociale, la psychologie du développement et la psychologie politique, ainsi que l'économie et les sciences de l'éducation. La plupart de ces protocoles repose sur la mesure des choix des participant·e·s entre différentes options ou sur l'évaluation de ces options sur une échelle. Pour cela, différents types de matériel (ou *stimuli*) peuvent être présentés aux participant·e·s (images, textes, vidéos, extraits sonores etc). Cette méthode permet ainsi de mesurer des attitudes et comportements de façon directe, ce qui peut être particulièrement utile lorsqu'il s'agit de comportements ou d'attitudes que les participant·e·s n'ont pas tendance à rapporter ou dont ils et elles n'ont pas conscience, même si ces attitudes peuvent influencer de façon non négligeable leurs comportements, comme c'est le cas pour les biais sexistes implicites.

En plus d'offrir la possibilité de mesurer de façon directe le comportement, l'expérimentation en laboratoire permet également de mesurer la façon dont un comportement peut être influencé par un contexte précis. Il s'agit du cœur de la méthode expérimentale scientifique : en comparant le comportement des participant·e·s dans

différentes conditions, une dans laquelle le facteur d'intérêt (celui dont on cherche à étudier l'influence) est présent et une dans laquelle il est absent, il est possible d'évaluer le lien de causalité entre ce facteur et le comportement étudié. Néanmoins, comme il s'agit d'études en laboratoire, ce facteur d'intérêt doit être extrait du contexte réel pour être étudié de façon expérimentale. Par exemple, dans le cadre de l'étude de l'acceptabilité d'un nouveau médicament, son prix, son efficacité et ses effets secondaires pourront être étudiés ensemble ou séparément à l'aide de choix fictifs afin d'estimer leurs influences sur les perceptions des participant·e·s. Ainsi, les expérimentations en laboratoire nécessitent d'effectuer en amont une analyse approfondie des facteurs pouvant affecter le comportement d'intérêt. Cette notion de comparaison s'étend au-delà des choix eux-mêmes et peut également être appliquée à différents contextes ou conditions. Par exemple, la comparaison entre une condition dans laquelle les participant·e·s ont accès à l'information sur le pourcentage d'élèves filles dans chaque filière d'enseignement secondaire à une condition et celle dans laquelle cette information n'est pas donnée permet d'estimer l'effet de ce type d'information sur les choix d'orientation des élèves.

Ces comparaisons peuvent être réalisées en présentant l'ensemble des contextes ou des choix à chacun des participant·e·s ou en ne présentant qu'un type de contexte ou de choix à chaque participant·e. La première méthode, appelée intra-participant·e·s, permet une estimation précise de ces effets en écartant la possibilité que les différences observées soient dues à d'autres facteurs que ceux manipulés dans l'expérience (comme des facteurs démographiques par exemple). En revanche, la seconde méthode, appelée inter-participant·e·s, ne permet pas de totalement exclure l'existence de variable explicative non-mesurée, mais est nécessaire lorsque les deux conditions manipulées sont incompatibles. Par exemple, dès que les participant·e·s ont reçu l'information sur le pourcentage d'élèves filles dans chaque filière, leurs choix seront très probablement influencés par ce facteur même si cette information n'est plus présente par la suite.

La mise en place et l'utilisation des expérimentations en laboratoire nécessitent donc plusieurs étapes de réflexion théorique, appuyées sur la connaissance de cette méthode mais aussi sur une analyse fine des politiques publiques, afin de garantir la qualité des données récoltées via le protocole expérimental (la validité interne de l'expérience) ainsi que leur capacité à expliquer des comportements et des situations pertinentes pour les politiques publiques (la validité externe de l'expérience).

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

La méthode d'expérimentation en laboratoire a une double utilité pour l'évaluation des politiques publiques. Tout d'abord elle offre un nouvel outil pour mesurer les comportements cibles des politiques (les comportements que les politiques cherchent à modifier), offrant des mesures complémentaires aux outils existants comme les questionnaires. Elle peut donc être intégrée dans le panel d'outils mobilisables *ex post* pour mesurer les changements de comportements à la suite de la mise en œuvre d'une intervention ou d'une politique publique.

Elle permet également de mieux connaître le comportement d'intérêt, d'en évaluer les composantes clés et d'éclairer l'élaboration des politiques publiques. Elle enrichit ainsi empiriquement la connaissance *ex ante* des comportements cibles pour permettre l'élaboration de politiques publiques mieux adaptées et ainsi potentiellement plus efficaces.

III. Exemples d'utilisation de cette méthode pour l'évaluation des politiques publiques dans les domaines de l'éducation, de la lutte contre les discriminations et de la propreté urbaine

Les méthodes d'expérimentation en laboratoire ont notamment été utilisées dans le domaine de l'éducation pour évaluer l'efficacité de différentes interventions, comme la pratique du sport, de la méditation ou du théâtre, sur les fonctions exécutives des enfants et des adolescent·e·s. Les fonctions exécutives sont un concept issu des sciences cognitives regroupant les processus psychologiques impliqués dans l'exécution d'action orientées vers un but, nécessitant entre autres le recours à la planification des actions, à l'inhibition de comportements concurrents et le passage de façon fluide d'une action à l'autre. Elles ont été montrées comme associées à plusieurs mesures de réussite scolaire, académique et professionnelle, ce qui a mené au développement d'interventions cherchant spécifiquement à les améliorer chez les enfants et les adolescents. Les fonctions exécutives étant mesurées de façon robuste par des expérimentations en laboratoire, comme des tâches lors desquelles les participant·e·s doivent inhiber une réponse automatique afin de fournir une réponse non-automatique, des expérimentations en laboratoire ont été utilisées pour évaluer l'efficacité de ces interventions. Ainsi, afin d'évaluer l'efficacité d'un programme de méditation de quatre semaines destiné à des élèves entre 9 et 11 ans, Parker et collaborateurs ont utilisé une tâche de Flanker, une tâche bien connue de mesure des fonctions exécutives, comparant les taux de bonnes réponses des participant·e·s dans différentes conditions : quand elles ou ils doivent indiquer l'orientation d'une image cible entourée d'autres images similaires et quand il leur est demandé d'indiquer uniquement l'orientation de ces autres images (Parker et col., 2014). Si cet exemple illustre la façon dont les concepts de sciences cognitives et les méthodes associées peuvent être mobilisés pour l'évaluation des politiques publiques, il est important de noter que ces méthodes peuvent être

combinées avec des outils issus d'autres champs comme les questionnaires. Ainsi, plusieurs interventions visant à réduire les biais racistes ont combiné des mesures de racisme explicite, obtenues à l'aide de questionnaires, et implicites, mesurées à l'aide d'expérimentations en laboratoire, afin d'obtenir une vision la plus complète possible des effets de ces interventions. À titre d'exemple, nous pouvons citer l'étude publiée par Devine et collaborateurs en 2012 qui a montré qu'une intervention combinant une explication sur l'existence de biais racistes implicites et la présentation de stratégies de réduction de ces biais conduite sur des étudiant·e·s américain·e·s n'avait pas d'effet significatif sur les biais implicite mais entraînait une réduction des biais racistes sur une durée de deux mois (Devine et col., 2012).

Les méthodes d'expérimentation en laboratoire ont également été appliquées pour évaluer *ex ante* les possibles effets de nouvelles politiques. Par exemple, en s'appuyant sur la littérature en politiques publiques sur l'importance de la visibilité des poubelles dans la réduction des déchets de rue, Abdel Sater et collaborateurs ont évalué en laboratoire, l'efficacité possible d'une intervention consistant à changer la couleur des sacs des poubelles de rue. Pour cela, ces chercheurs ont comparé la capacité des participant·e·s à détecter des poubelles dans des photos de rue en fonction de la couleur des sacs poubelle. Celle-ci avait été manipulée par ordinateur à partir de photos réelles, pour que la tâche expérimentale soit la plus proche possible de conditions réelles mais également la plus contrôlée possible : seule la couleur des sacs poubelles différenciait les photographies avec les sacs de couleurs gris et de celles avec les sacs de couleur rouge. Cette étude a permis de mettre en évidence l'efficacité potentielle de cette intervention simple et à faible coût sur la visibilité des poubelles (Abdel Sater et col., 2020). Bien que cet exemple n'ait pas encore été traduit en l'implémentation d'une intervention réelle, il illustre la façon dont les expérimentations en laboratoire peuvent être intégrées au cycle d'élaboration des politiques publiques.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Que son utilisation soit *ex post* ou *ex ante*, le premier élément à prendre en compte pour évaluer de la pertinence de l'utilisation de l'expérimentation en laboratoire pour l'évaluation des politiques publiques est l'alignement entre le comportement d'intérêt, celui qui est directement lié à la question de politiques publiques, et le comportement mesuré en laboratoire. Cette idée est fondamentale pour que les expérimentations en laboratoire soient réellement utiles pour l'évaluation des politiques publiques et ne soient pas un simple outil marketing. En effet, l'expérimentation en laboratoire utilise parfois des tâches abstraites, souvent construites initialement pour évaluer des mécanismes psychologiques fondamentaux tels que la motivation. Il faut donc s'assurer que le comportement mesuré expérimentalement est bien associé de façon robuste aux comportements d'intérêt tel qu'il est observé dans les situations réelles. Cette question est d'autant plus importante que les expérimentations en laboratoire permettent de mesurer non seulement des attitudes explicites, celles que les participant-e-s sont prêts à rapporter lors d'entretiens ou d'enquêtes, mais aussi des attitudes implicites, dont les participant-e-s eux-mêmes et elles-mêmes n'ont pas forcément conscience. Si ce dernier type d'attitude présente un fort intérêt théorique, il est seulement faiblement prédictif des comportements des individus dans la vie courante et ne permet de prédire le comportement que dans des situations précises, comme quand les individus doivent prendre une décision extrêmement rapidement. Ainsi, une intervention peut ne pas avoir d'effet significatif sur les attitudes implicites mais tout de même modifier le comportement des participant-e-s. Ces deux niveaux de mesure peuvent avoir un intérêt pour l'évaluation en profondeur des politiques publiques et l'anticipation de potentiels effets non prédits mais n'utiliser qu'une mesure au niveau implicite pour l'évaluation des politiques publiques peut en revanche entraîner des interprétations erronées de leur efficacité.

D'autre part, il est important, comme pour tout outil mobilisé dans le cadre de l'évaluation des politiques publiques, de considérer la taille des effets obtenus. En effet, le contexte artificiel dans lequel sont observés les effets dans le cadre des expérimentations en laboratoire appelle à la prudence lors de la mobilisation de ces résultats pour l'évaluation des politiques publiques. Ces conditions et tâches souvent très artificielles, bien qu'elles permettent d'isoler le plus possible le comportement et les facteurs d'intérêt, peuvent également mener à des interprétations biaisées au moment de généraliser ces résultats à des situations réelles. En effet, une expérimentation dans laquelle un seul type d'information est donné (par exemple, le nom du journal ayant diffusé un article), peut mener à surestimer le poids de ce type d'information dans les décisions des individus, car contrairement au contexte expérimental, en contexte réel les individus peuvent fonder leurs choix sur une multitude d'informations disponibles. La mobilisation de l'expérimentation en laboratoire pour l'évaluation des politiques publiques nécessite donc de prendre en compte le protocole expérimental utilisé dans son ensemble, c'est-à-dire non seulement le type de choix qui a été mesuré, mais également le type d'informations auxquelles les participant·e·s avaient accès.

Enfin, s'il s'agit d'une utilisation *ex ante*, il est également important de prendre en compte la population sur laquelle les résultats ont été obtenus afin d'évaluer si ces résultats sont mobilisables pour la population cible de la politique publique. En effet, des résultats comportementaux obtenus uniquement sur une population particulière peuvent ne pas être valides dans une autre population. Ces différences entre populations sont également importantes lorsque l'analyse porte spécifiquement sur la comparaison entre populations ou consiste en l'utilisation d'un protocole expérimental pour une population différente de celle sur laquelle il a initialement été testé. Dans ces deux cas, il est nécessaire de considérer que les variations de comportement observées expérimentalement peuvent être dues à la structure du protocole expérimental lui-même et non à des différences dans le comportement d'intérêt. Par exemple,

des différences sur le niveau de concentration des participant·e·s sur la tâche à réaliser en laboratoire peuvent générer des différences de comportement qui ne reflètent pas des différences réelles dans le comportement cible. Il est donc crucial que le type d'expérience choisi soit cohérent avec la ou les populations cibles afin de ne pas créer artificiellement des différences de comportement entre populations ou de ne pas sous-estimer ou surestimer l'existence de certains comportements dans ces populations.

Enfin, à ces éléments directement liés à la mobilisation des expérimentations en laboratoire à l'évaluation des politiques publiques s'ajoutent les critères généraux d'évaluation de la qualité des expérimentations en laboratoire. Ces critères reposent notamment sur l'évaluation de la sensibilité de l'expérimentation et de ses résultats à l'influence de biais comportementaux et à l'aléatoire. Pour cela, l'utilisation de formulations spécifiques, la répétition des questions, l'utilisation d'une diversité de matériel expérimental contrôlé sur des éléments clés (comme l'utilisation d'une série de visages de femmes et d'hommes différents mais ayant la même expression pour évaluer les biais sexistes) et la présentation de façon aléatoire des différents éléments de l'expérience (l'ordre de présentation des questions et des conditions par exemple) sont classiquement mis en place pour s'assurer de la fiabilité des résultats des expérimentations conduites en laboratoire.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Les deux principaux atouts de la méthode expérimentale en laboratoire sont d'une part, de permettre de tester l'existence de liens de causalité entre un facteur ou un contexte et un comportement et, d'autre part, d'offrir un outil de mesure spécifique des comportements et des attitudes. Toutefois, il est important de noter que les critères nécessaires

à la réalisation d'une expérience en laboratoire fiable rendent cette méthode parfois plus contraignante que d'autres méthodes. Par exemple, les expérimentations en laboratoire sont souvent plus longues que les enquêtes par questionnaires, rendant cette méthode plus coûteuse. Parallèlement, la nécessité de contrôler un grand nombre de facteurs limite le caractère exploratoire de cette méthode et la rend plus appropriée à la mesure d'un comportement précis ou l'évaluation d'une hypothèse donnée.

De plus, le contexte très contrôlé des expériences en laboratoire limite la possibilité d'interpréter directement les résultats de ces expériences en termes de comportements en dehors du laboratoire. Les comportements d'intérêt sont d'ailleurs parfois mieux prédits par des réponses explicites que par des mesures réalisées lors d'expérimentation en laboratoire. Toutefois, les expérimentations en laboratoire permettent de mesurer des comportements qui sont difficiles voire impossibles à mettre en évidence lors d'entretiens ou à mesurer dans des enquêtes classiques. En effet, celles-ci offrent la possibilité de mesurer comportements implicites et sont moins sensibles aux biais récurrents observés avec les autres méthodes, notamment le biais de désirabilité sociale, c'est-à-dire la volonté des participant·e·s de se montrer sous leur meilleur jour et de répondre selon ce qu'ils ou elles perçoivent être une norme sociale, bien que cela reste tout de même un risque dans les expérimentations en laboratoire. Ainsi, les expérimentations en laboratoire sont particulièrement prometteuses pour l'évaluation de l'efficacité des politiques publiques à modifier non seulement les comportements des individus mais également les biais implicites pouvant avoir des effets importants à long terme.

Références bibliographiques citées dans le texte

Abdel Sater, Rita. et Mus, Mathilde. et Wyart, Valentin. et Chevallier, Coralie. 2020. A zero-cost attention-based approach to promote cleaner streets.

Devine, Patricia. et Forscher, Patrick. et Austin, Anthony. et Cox, William. 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, 48(6): 1267-1278.

Parker, Alison. et Kupersmidt, Janis. et Mathis, Erin. et Scull, Tracy. et Sims, Calvin. 2014. The impact of mindfulness education on elementary school students: Evaluation of the Master Mind program. *Advances in School Mental Health Promotion*, 7(3): 184-204.

Quelques références bibliographiques pour aller plus loin

Bordens, Kenneth. et Abbott, Bruce. 2014. *Research Design and Methods: A Process Approach*. McGraw Hill.

Gawronski, Bertram. 2009. Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology/Psychologie canadienne*, 50(3): 141-150.

Reis, Harry. et Judd, Charles. 2000. *Handbook of research methods in social and personality psychology*. Cambridge University Press.

7. Testing

NICOLAS JACQUEMET

Résumé

Le *testing*, ou méthode d'étude par correspondance, est une méthode quantitative permettant de mesurer les discriminations. Elle passe par l'envoi de candidatures fictives en réponse à des offres réelles (par exemple, offres d'emploi). Par l'objectivation des comportements discriminatoires, cette méthode est très utile, sur le plan prospectif, pour la conception des politiques antidiscriminatoires.

Mots-clés : Méthodes quantitatives, correspondance, discriminations, politiques antidiscriminatoires, candidatures expérimentales

I. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Les discriminations correspondent à une inégalité de traitement sur la base de caractéristiques individuelles qui ne devraient pas être pertinentes pour la décision à prendre : favoriser une candidature masculine dont les compétences professionnelles sont supérieures n'est pas discriminatoire; mais écarter une candidature féminine sur la base du soupçon que la disponibilité de cette candidate sera moindre que celle d'un candidat dont le profil est équivalent est bel et bien discriminatoire, car rien ne permet de penser que la candidate en question correspond à ce stéréotype. À ce titre, les discriminations sont une source d'inégalité majeure. Plus encore que d'autres types d'inégalités, elles sont à la fois très coûteuses sur le plan économique, en privant l'économie de certains

de ses talents; et persistantes, car l'anticipation de telles inégalités de traitement découragent les personnes discriminées et les conduisent à faire des choix (de niveau et de filière d'éducation, de parcours professionnel) qui ne font qu'amplifier ces inégalités de départ.

Malgré l'importance des enjeux, l'élaboration de politiques publiques destinées à lutter contre les discriminations souffre du manque d'éléments de diagnostic dû à la très grande difficulté de la mesurer. C'est l'objectif poursuivi par la méthode d'étude par correspondance (*testing*).

II. En quoi consiste cette méthode?

Bien que cette méthode puisse être appliquée à de très nombreux secteurs de l'activité économique (recherche de logement, locations saisonnières, candidature des étudiants en master) et à de nombreuses sources différentes de discrimination (la religion, les préférences sexuelles, les origines socio-économiques, le lieu de résidence, le handicap) cette présentation se concentre par simplicité sur l'application de cette méthode à la mesure des discriminations à l'embauche fondées sur le sexe et / ou l'origine.

Cette méthode quantitative est conçue de manière à fournir une mesure du succès de différents types de candidats¹ en fonction de leurs caractéristiques socio-démographiques, mais tout en neutralisant l'effet de la qualité intrinsèque des candidatures. Chacun de ces deux objectifs a des implications méthodologiques qui lui sont propres.

1. L'auteur ne souhaite pas utiliser l'écriture inclusive.

Constitution des candidatures fictives

Le succès de différents types de candidat-e-s est observé grâce au recours à des candidatures artificielles, envoyées en réponse à des offres d'emploi réelles circulant sur le marché du travail. La méthode combine pour ce faire trois ingrédients : des identités, des candidatures, et des offres d'emploi.

Les caractéristiques socio-démographiques dont on cherche à mesurer l'effet sont véhiculées par l'identité du ou de la candidat-e. Pour tester à la fois les discriminations liées au sexe et les discriminations qui affectent les candidat-e-s issu-e-s, par exemple, de l'immigration maghrébine, il s'agira donc de créer une liste de quatre identités fictives (ou quatre catégories d'identités différentes) : deux patronymes à consonnance française, l'un associé à un prénom masculin et l'autre à un prénom féminin, et deux patronymes qui suggèrent que la personne est issue de l'immigration maghrébine, associées aux mêmes variations du prénom. Chacune de ces identités se voit dotée d'un numéro de téléphone unique et d'une adresse électronique permettant d'entrer en contact avec les candidat-e-s. Ces noms et prénoms ainsi que ces informations de contact correspondent au bloc identité des candidatures.

Pour répondre à des offres d'emploi, ces identités sont portées sur des candidatures qui combinent le plus souvent un CV et une lettre de motivation. L'objectif est de construire des candidatures qui soient aussi crédibles que possible, et permettront de distinguer le succès de différentes identités. Il faut donc que le processus de construction des candidatures conduise à une qualité qui ne soit ni trop élevée ni trop faible en comparaison des candidatures réelles qui seront reçues, car toute candidature qui conduit à un traitement indifférencié des candidats, qu'il soit positif ou négatif, rend impossible l'identification des caractéristiques favorisant le succès des candidatures expérimentales.

La construction du CV nécessite de choisir le contenu de la section formation ainsi qu'une liste d'expériences qui soient toutes deux réalistes et compatibles avec le métier pour lequel la candidature est envoyée, ainsi qu'une section dédiée aux activités extra-professionnelles. Afin de garantir la correspondance entre ces éléments de CV et les métiers visés, la plupart des études collectent de vrais CV (disponibles, par exemple, en ligne), mélangent les informations issues de plusieurs CV pour construire un CV unique puis modifient à la marge les sections « expérience », « formation », et « activités extra-professionnelles » ainsi obtenues. Le contenu du CV est complété par un bloc contenant des informations personnelles permettant d'indiquer *a minima* l'adresse postale, à laquelle peuvent éventuellement être associés le statut marital, la présence d'enfants, l'âge ou encore la date de naissance. La mise en forme de ces informations nécessite de choisir autant de modèles prédéfinis que de CV différents, qui détermineront l'ordre des sections, la police de caractère utilisée ainsi que l'organisation des différentes informations (de nombreux modèles dans différents formats de fichier peuvent être facilement trouvés en ligne). Les lettres de motivation sont construites de la même manière, en combinant le contenu de lettres de motivation existantes. Les accords de genre seront adaptés en fonction du sexe de l'identité portée sur la candidature (si le sexe fait partie des caractéristiques testées, il est souhaitable de choisir les formulations qui permettent de multiplier les accords de genre). Le couple formé par un CV et une lettre de motivation correspond à une candidature.

Les offres d'emploi auxquelles ces candidatures seront envoyées sont collectées sur des sites publics d'information (il s'agit souvent du site du Pôle emploi, qui est le service public de l'emploi en France), mais selon les métiers visés il est parfois nécessaire de recourir à des sites spécialisés). Ces offres sont filtrées pour vérifier qu'elles correspondent aux critères d'inclusion prédéfinis, qui concernent au premier chef le métier et la localisation de l'emploi, mais aussi par exemple l'exigence d'expériences ou de compétences spécifiques. Les offres pour lesquelles il n'est pas possible d'envoyer une candidature selon les modalités préalablement

définies (souvent, par mail, mais aussi lorsque par exemple le dépôt d'une candidature nécessite de compléter un questionnaire en ligne) sont systématiquement écartées. Pour les offres d'emploi restantes, qui seront incluses dans l'étude, l'ensemble des caractéristiques de l'offre (durée, type de contrat, salaire, etc.) sont soigneusement enregistrées afin de constituer une base permettant de documenter l'hétérogénéité observée des offres d'emploi.

Le nombre de candidatures expérimentales qui seront envoyées en réponse à une offre d'emploi donnée (qui va de pair avec le nombre de candidatures différentes qu'il est nécessaire de construire) est un choix délicat. D'un point de vue statistique, il est très avantageux de pouvoir comparer le succès de différentes candidatures en réponse à une offre d'emploi donnée (i.e., comparaisons « intra-offre »), car une telle comparaison élimine l'effet de tous les éléments inobservés qui sont spécifiques à l'offre d'emploi et améliore donc la précision statistique des mesures. L'envoi de plusieurs candidatures correspondant à un groupe socio-démographique donné permet également de mesurer plus finement les caractéristiques de la distribution de la discrimination entre les différentes annonces (voir les résultats présentés dans Kline et al. 2020). S'il est par conséquent souhaitable d'envoyer plusieurs candidatures en réponse à chaque offre, le nombre maximum est limité par deux facteurs. D'une part, la multiplication des candidatures augmente les perturbations induites par la réalisation de l'étude sur le fonctionnement du marché du travail et, surtout, le risque de détection. Ce risque peut être contenu en prenant soin de laisser un délai suffisant entre l'envoi de deux candidatures, mais ce délai va de pair avec une diminution de la probabilité de succès pour les candidatures les plus tardives et ce d'autant plus que le métier attire un nombre important de candidatures. D'autre part, certains travaux récents (Philips, 2019) montrent que le portefeuille de candidatures envoyées en réponse à une offre donnée est susceptible d'affecter le succès relatif des candidatures expérimentales. L'augmentation du nombre de candidatures augmente les risques de tels biais.

Ces deux facteurs conduisent à être d'autant plus restrictif quant au nombre de candidatures envoyées que le métier est en tension. La combinaison de ces différents facteurs conduit la plupart des études à se limiter à l'envoi de quatre candidatures au maximum en réponse à chaque offre d'emploi, envoyées jusqu'à 24h au plus tard après leur publication. Pour ce faire, chaque identité est associée à une candidature unique (un CV et une lettre de motivation), conduisant à autant de candidatures expérimentales uniques et distinctes que le nombre d'envois en réponse à chaque offre.

La mesure du succès des candidatures expérimentales nécessite de conserver une trace précise, et limitée dans le temps (en ignorant par exemple les réponses reçues plus de 3 mois après leur envoi), de la communication des employeurs avec les candidats en archivant toute correspondance écrite et en retranscrivant le contenu des messages téléphoniques reçus. Ces réponses sont ensuite classifiées afin de distinguer les refus, les non-réponses, les demandes d'informations complémentaires et les convocations à un entretien d'embauche (qui sont parfois regroupées sous le terme de 'manifestations d'intérêt'). Pour des raisons éthiques évidentes, il est impératif de décliner toute manifestation d'intérêt aussi vite que possible, préférentiellement selon les mêmes modalités de contact et en suivant un script prédéfini (qui prétexte le plus souvent l'acceptation antérieure d'une offre d'embauche).

Protocole d'assemblage

La combinaison de l'ensemble de ces ingrédients fournit une mesure du succès de candidatures fictives qui se distinguent, entre autres, par le groupe socio-démographique auquel est associé l'identité portée sur la candidature. Bien évidemment, de telles différences de succès peuvent également être liées au contenu de la candidature elle-même, ce qui est d'autant plus probable que les candidatures sont nettement différentes

les unes des autres. Une solution pourrait donc être de s'assurer que les candidatures sont aussi proches que possible les unes des autres. Mais outre que toute différence, même infime, entre les candidatures conduit à la même conclusion, il est particulièrement difficile de distinguer des différences négligeables de différences plus importantes, car les différences qui sont pertinentes concernent les variations subjectives de la qualité des candidatures telle que perçues par les personnes en charge du recrutement.

Le protocole qui permet aux études par correspondance de neutraliser l'effet de toutes les caractéristiques des candidatures expérimentales qui pourraient être confondantes (i.e., dont l'impact sur le taux de succès conduirait à des conclusions erronées quant à l'effet des caractéristiques d'intérêt) consiste à organiser une rotation systématique de l'association entre identités d'une part et candidatures d'autre part. Si, par exemple, l'identité A est portée sur la candidature A et l'identité B sur la candidature B lors du premier envoi, ces associations seront inversées lors de l'envoi suivant (l'identité A apparaissant sur la candidature B) avant de revenir à l'association initiale lors du troisième envoi, etc. Cette rotation n'élimine pas l'effet de la qualité perçue de la candidature : si la candidature A se trouve être de meilleure qualité, le succès de l'identité qui lui est associée en sera affecté. Mais la rotation permet de s'assurer que toute différence systématique de succès associée à l'identité sur l'ensemble des envois ne puisse plus être attribuée au contenu de la candidature. D'un point de vue statistique, toute caractéristique pour laquelle une rotation systématique est organisée devient une source de bruit dans la mesure de la discrimination liée aux caractéristiques d'intérêt, c'est à dire une source de variation du taux de succès qui distingue les candidatures appartenant à différentes catégories sans toutefois être imputables à la discrimination. Par construction, ce bruit est indépendant des caractéristiques dont on cherche à mesurer l'effet et ne nuit donc pas à la capacité de la méthode de mesurer la discrimination, mais rend sa détection plus difficile. Ces conséquences du bruit dans les mesures peuvent être réduites en

adaptant l'analyse statistique en conséquence (sous la forme d'effets fixes offres), mais une telle modélisation suppose un effet homogène de la qualité des candidatures sur l'ensemble des employeurs.

Au total, la méthode d'étude par correspondance repose donc sur trois principes : démultiplier le nombre de candidatures expérimentales afin de mesurer l'effet des caractéristiques socio-démographiques par lesquelles elles se distinguent, s'assurer de la plus grande homogénéité possible de ces candidatures afin de réduire le bruit qui affectera la mesure de leur effet, et organiser une rotation systématique de l'association entre les profils socio-démographiques et toute autre caractéristique susceptible d'en affecter le succès. Ces trois principes constituent une boîte à outils qui peut être déclinée à de très nombreux aspects du fonctionnement du marché du travail. On peut ainsi mesurer l'effet, par exemple, de la présence d'épisodes de chômage dans le parcours professionnel en modifiant expérimentalement la section « expérience » des candidatures, de la distance entre le domicile et le travail ou du lieu de résidence en manipulant l'adresse d'habitation, ou de la situation familiale en faisant varier le bloc identité selon la présence d'enfants ou le statut marital.

III. Un exemple d'utilisation de cette méthode

Une étude récente réalisée conjointement par l'Institut des Politiques Publiques et ISM Corum sous l'égide de la DARES est l'une des premières études de grande ampleur permettant de dresser un panorama des inégalités d'accès à l'emploi selon le sexe et l'origine sur le marché du travail français (Dares IPP et ISM Corum, 2021a et 2021b). Ces résultats confirment l'existence de discriminations liées à l'origine qui sont à la fois fortes et transversales à l'ensemble des métiers étudiés, conduisant à un handicap de l'ordre 30% dans les chances de recevoir une réponse positive. Cette étude met également en évidence l'absence de discrimination liée au sexe du candidat, suggérant que, contrairement à

une idée reçue persistante, les fortes inégalités de carrière qui existent sur le marché du travail entre les hommes et les femmes ne peuvent pas être attribuées aux décisions d'embauche.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Le niveau du taux de rappel pour un type de candidature donné ne fournit que peu d'informations sur le fonctionnement du marché du travail. Les résultats issus de la méthode du *testing* proviennent des comparaisons de taux de rappel entre différents types de candidatures. Pour que ces comparaisons permettent de détecter l'écart de succès que rencontrent différents types de candidats, il est important que les taux de rappel des candidatures de référence soient suffisamment élevés.

Les variations des caractéristiques socio-démographiques sont introduites par l'intermédiaire de l'identité portée sur les candidatures, qui est supposée affecter les perceptions des personnes en charge du recrutement. Pour s'en assurer, il est de plus en plus fréquent dans les études de *testing* de faire précéder l'étude d'une enquête dans laquelle un échantillon de répondants doit associer un sexe et/ou une origine à chacune des identités qui leur sont présentées. Cette enquête fournit une mesure empirique de la qualité des perceptions induites par les identités, et peut permettre de sélectionner les identités en conservant celles dont les perceptions sont les plus cohérentes avec le groupe d'appartenance souhaité. Une telle enquête peut également être l'occasion de collecter des informations supplémentaires sur le profil perçu des identités présentées : des travaux récents montrent en effet que les identités véhiculent de nombreux stéréotypes liés par exemple à la classe sociale ou à la zone d'habitation qui peuvent contribuer aux différences de succès observées des candidatures de différentes catégories (Gaddis, 2017).

Enfin, les différences observées de taux de rappel sont sujettes à la fameuse critique connue sous le nom de « critique d'Heckman », selon laquelle des différences perçues par les employeurs et employeuses dans la variance des compétences à l'intérieur des différents groupes de population suffirait à produire des écarts systématiques en moyenne dans les taux de rappel, et serait interprétée à tort comme un biais systématique à l'encontre de ces groupes de population. L'existence de différences suffisantes de qualité entre les différentes candidatures expérimentales peut permettre de prendre en compte cet effet en menant des analyses statistiques qui autorisent des variances dans les éléments inobservés qui diffèrent entre groupes (Neumark, 2012).

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Le protocole de mesure fourni par les *testing* permet d'évaluer de manière précise et convaincante l'ampleur des pratiques discriminatoires et l'effet spécifique des caractéristiques socio-démographiques des candidats sur le succès de leur insertion sur le marché du travail. Il permet d'objectiver un phénomène que les approches qualitatives ont plus de difficultés à révéler : ces pratiques ne sont pas aisément verbalisées en entretien semi directif par exemple, car illégitimes et parfois non conscientes. Son principal atout est de garantir par construction l'indépendance entre ces caractéristiques et l'ensemble des autres éléments qui composent la candidature. La principale alternative est d'utiliser des données d'enquête pour étudier les différentiels de parcours sur le marché du travail entre différentes catégories de population. Mais de telles études nécessitent des hypothèses statistiques fortes, et souvent peu crédibles, destinées à neutraliser l'effet des différences d'éducation ou de parcours professionnel qui distinguent ces groupes de population et contribuent aux écarts de succès observés sur le marché du travail.

La portée des résultats produits par l'application de cette méthode est néanmoins limitée par deux facteurs importants.

Le premier est que la mesure du succès des candidatures repose entièrement sur la convocation, ou non, à un entretien d'embauche. Or cette étape n'est que le reflet imparfait du résultat final du processus de recrutement : l'existence d'une discrimination à cette étape du processus ne prédit une discrimination lors de l'embauche effective qu'à condition que tous les candidats convoqués soient traités sur un pied d'égalité. Si, au contraire, une discrimination supplémentaire s'exerce lors des entretiens d'embauche, les mesures de discrimination fournies par cette méthode sous-estiment le phénomène. S'il s'avère enfin que les populations discriminées dans le tri des candidatures sont favorisées dans des proportions strictement inverses lors du choix du candidat final, alors ces mesures distordent la réalité des discriminations. Les méthodes d'audit, qui consistent à faire appel à des acteurs jouant le rôle de candidats expérimentaux mais réels, permettent de dépasser cette limite, mais elles présentent l'inconvénient de faire intervenir un ensemble très large de facteurs (l'apparence physique, la voix) qui sont susceptibles d'influencer le processus de recrutement mais ne peuvent pas être distingués des caractéristiques socio-démographiques qui sont apparentes.

La seconde limite est commune à toute étude empirique mais s'avère particulièrement aiguë dans le cas des *testings* : comme discuté plus haut, les mesures sont d'autant plus précises que les candidatures sont homogènes. S'y ajoutent des raisons pratiques, liées au fait que le nombre et la spécificité des candidatures fictives augmentent avec la diversité (géographique ou en termes de métiers) des offres d'emploi. En conséquence, les études de *testing* sont souvent circonscrites à un périmètre limité, et leurs résultats ne peuvent donc qu'être conditionnels au champ d'application de l'étude en termes de type d'emploi, de secteur d'activité, de zone géographique, de tranche d'âge des candidats, etc. La généralisation à l'ensemble du marché du travail des résultats observés dans le cadre de ce type d'étude repose donc sur l'hypothèse que le

périmètre choisi ne présente pas de spécificités en termes de propension à discriminer (préférences des recruteurs, degré de concurrence dans le recrutement, etc.) ou, de façon plus convaincante, sur l'accumulation d'études concordantes portant sur différentes sphères du marché du travail.

Quelques références bibliographiques pour aller plus loin

Adamovic, Mladen. 2020. « Analyzing Discrimination in Recruitment: A Guide and Best Practices for Resume Studies ». *International Journal of Selection and Assessment* 28, n°4 (2020): 445-64.

Adida, C.L.. et Laitin, D.D.. et Valfort, M.-A.. 2010. Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences* 107, 22384-22390.

Dares IPP et ISM Corum. 2021a. « Discrimination à l'embauche selon le sexe: les enseignements d'un testing de grande ampleur ». Dares Analyses n°26/Note IPP n°67.

Dares IPP et ISM Corum. 2021b. « Discrimination à l'embauche des personnes d'origine supposée maghrébine: quels enseignements d'une grande étude par testing? ». Note IPP n°76/Dares Analyses n°67.

du Parquet, Loïc. et Petit, Pascale. 2019. « Discrimination à l'embauche: retour sur deux décennies de testings en France ». *Revue française d'économie* Vol. XXXIV, n°1: 91-132.

Edo, Anthony. et Jacquemet, Nicolas. 2013. La discrimination à l'embauche sur le marché du travail français. Opuscule du CEPREMAP n°31, Éditions rue d'Ulm.

- Fougère, Denis. et Rathelot, Roland. et Aeberhardt, Romain. « Commentaire: Les méthodes de testing permettent-elles d'identifier et de mesurer l'ampleur des discriminations? ». *Économie et Statistique* 447, n°1 (2011): 97-101.
- Gaddis, S Michael. 2017. « How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies ». *Sociological Science* 4: 469-89. <https://doi.org/10.15195/v4.a19>.
- Kline, Patrick M.. et Walters, Christopher R.. 2020. « Reasonable doubt: Experimental detection of job-level employment discrimination ». *Econometrica* 89, n°2 (2020): 765-92.
- Neumark, David. 2012. « Detecting Discrimination in Audit and Correspondence Studies ». *Journal of Human Resources* 47, n°4: 1128-57.
- Phillips, D.C.. 2019. Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments. *Economic Journal* 129, 2240–2264.

8. L'analyse coût-efficacité

THOMAS RAPP

Résumé

L'analyse coût-efficacité est une méthode quantitative consistant à comparer le « retour sur investissement » d'une politique donnée (les résultats souhaités qu'elle produit, rapportés à son coût), par comparaison avec d'autres politiques possibles. Cette méthode permet d'estimer l'efficacité d'une politique, c'est-à-dire sa capacité à maximiser un critère de résultats pour chaque euro de dépense publique. Elle est utile pour guider les choix de politique publique et d'allocation des dépenses publiques au sein d'un secteur donné.

Mots-clés : Méthodes quantitatives, coût/efficacité, efficience, efficacité

I. En quoi consiste cette méthode?

L'analyse coût-efficacité est une méthode qui consiste à explorer l'efficience d'une politique publique, c'est-à-dire dans le langage familier à déterminer son « retour sur investissement ». C'est une méthode comparative dans laquelle l'intervention évaluée est comparée à plusieurs autres options : politiques existantes, alternatives, etc. Cette comparaison permet de hiérarchiser les différentes options et de déterminer laquelle permet d'optimiser la dépense publique, c'est-à-dire d'obtenir le meilleur résultat possible pour chaque euro investi. La hiérarchisation des différentes options est réalisée à partir d'un calcul économique *a priori*

simple, celui du « ratio différentiel coût efficacité ». Ce calcul met en relation la différence de coûts et la différence d'efficacité constatée entre l'intervention et ses comparateurs.

Cinq étapes sont nécessaires à la mise en œuvre de cette méthode d'évaluation.

En premier lieu, il est nécessaire de choisir un point de vue pour l'analyse et une population cible. Cela consiste à déterminer quelle est la perspective adoptée pour le calcul : celle du financeur ou de la financeuse de la politique publique (payeur·euse)? de ses bénéficiaires? de la société dans son ensemble? On notera que la perspective souvent choisie dans les analyses coût-efficacité est la perspective sociétale, parce qu'elle prend en compte l'impact de la politique pour l'ensemble des parties prenantes (payeur·euse·s, bénéficiaires, etc.). Néanmoins, si cette perspective collective est plus intéressante pour le décideur public ou la décideuse publique, elle est aussi plus difficile à mettre en œuvre parce qu'elle implique une mesure exhaustive des coûts et des critères d'efficacité. Le choix de la population cible est souvent dicté par l'objectif de la politique publique. Par exemple, une campagne de prévention publique du cancer du sein ciblant des âges précis. Il est souvent pertinent de déterminer au sein de cette population des sous-groupes de personnes spécifiques en fonction par exemple de leur accès à l'intervention (par exemple, l'accès aux centres de soins) ou de leur exposition à d'autres mesures (par exemple, l'accès à un soin préventif financé par l'assurance privée).

En deuxième lieu, il faut déterminer le périmètre des coûts associés à la politique publique évaluée. Les coûts considérés dans l'évaluation sont ceux liés au déploiement de l'intervention, aussi appelés « coûts directs » : investissement d'infrastructures, de matériel, salaires des personnels dédiés à l'intervention, campagnes d'information, etc. Dans le cadre d'une évaluation portant sur plusieurs années, ces coûts sont actualisés pour tenir compte des variations d'inflation. Les coûts indirectement associés à la politique publique peuvent également être intégrés dans l'évaluation. En effet, si la politique publique a un impact fort sur la sphère

domestique, il est souvent judicieux d'inclure cet impact dans le calcul du ratio d'efficience. Par exemple, on peut s'attendre à ce qu'un accroissement de la générosité des aides publiques pour l'autonomie des seniors fragiles réduise l'absentéisme au travail des aidants familiaux, qui peuvent substituer à leur temps d'aide des services professionnels.

La troisième étape consiste à choisir un critère d'efficacité de la politique publique. Le choix de ce critère est déterminant, car il doit être assez « sensible » pour saisir l'impact de la politique. Deux grandes catégories de critères sont généralement mobilisées dans les évaluations : des critères de résultats et des critères d'utilité. D'une part, les critères de résultats permettent de mesurer l'efficacité de la politique évaluée avec des indicateurs chiffrés : par exemple des durées de chômage pour l'évaluation de politiques d'emploi, des taux de mortalité pour des politiques de santé, des taux de réussite scolaire pour des politiques d'éducation, etc. D'autre part, les critères d'utilité permettent de mesurer l'impact de la politique publique sur le bien-être des ménages, mesuré à l'aide de questionnaires. On parle alors d'une analyse coût-utilité, dont l'objectif est de mesurer si la politique évaluée permet d'optimiser le niveau moyen de bien-être de ses bénéficiaires.

En quatrième lieu, il faut déterminer quelles sont les sources de données à mobiliser pour mener l'évaluation, et estimer l'impact de la politique publique à partir de ces données. Les évaluations utilisent deux principales catégories de bases de données : celles issues « d'expérimentations aléatoires » et les données dites « de vie réelle ». Les expérimentations aléatoires (voir fiche séparée) consistent à comparer deux populations différentes, l'une recevant l'intervention évaluée et l'autre bénéficiant d'une alternative. Le succès de ces expérimentations repose sur l'absence de « contamination » entre les deux groupes comparés. En effet, toute interaction entre les membres des deux groupes remet en cause l'évaluation de l'efficacité du traitement, et donc toute l'évaluation coût-efficacité. Généralement, ces études portent sur une période de deux ans et incluent quelques milliers de participant·e·s. Les évaluations réalisées à partir de données de vie réelle consistent à

identifier *ex post* dans les bases de données administratives les deux populations incluses dans l'évaluation (toujours en fonction de leur exposition), et de suivre leur évolution pendant une période plus longue. Elles présentent l'avantage de leur exhaustivité : taille d'échantillon, nombre d'années de suivi, etc.

Enfin, on peut procéder à l'analyse coût-efficacité en elle-même. Cette analyse comporte deux étapes : il faut, d'une part, estimer les effets moyens observés sur les paramètres de coût et d'efficacité à partir des données et, d'autre part, modéliser l'impact coût-efficacité de la politique publique à partir des effets estimés. L'estimation des effets est réalisée à partir de régressions économétriques, qui permettent d'identifier l'impact moyen de la politique publique dans l'échantillon de données utilisé, et des intervalles de confiance pour cet effet (bornes supérieures et inférieures qui encadrent sa valeur réelle). Cette étape d'estimation est essentielle, car elle permet de définir si l'impact de la politique est significatif ou pas, autrement dit, si la politique donne en moyenne les résultats escomptés. Le choix de la méthode d'estimation est un aspect essentiel de cette étape. Ensuite, intervient la phase de modélisation, au cours de laquelle les estimations des effets obtenues dans la première étape vont être utilisées comme *inputs* pour le modèle d'évaluation coût-efficacité. Par exemple, on estimera différentes probabilités d'hospitalisations en urgence dans le cadre de l'évaluation d'une politique de prévention des chutes chez les seniors âgé·e·s. On testera alors tous les impacts possibles de la politique publique pour tenir compte de l'incertitude liée à ses effets, c'est-à-dire des différentes valeurs possibles pour les probabilités d'hospitalisation. On parle alors de « simulation ».

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

L'évaluation coût-efficacité permet d'apporter un éclairage sur l'efficacité d'une politique publique, c'est-à-dire sa capacité à maximiser un critère de résultat pour chaque euro de dépense publique. C'est donc un outil d'aide à la décision qui permet de répondre à de très nombreuses questions évaluatives : Est-il plus rentable de mettre en œuvre une politique A par rapport à une politique B? Quels sont les coûts et les bénéfices incrémentaux liés à l'adoption d'une politique publique? Quelles populations spécifiques peuvent bénéficier le plus du déploiement de cette politique publique? Comment améliorer l'allocation des dépenses publiques dans un secteur donné (santé, éducation, sécurité, etc.)? Les effets attendus d'une politique publique dépassent-ils les coûts de sa mise en œuvre? L'efficacité d'une politique peut-elle varier en fonction du profil de ses bénéficiaires, de la population couverte?

La réponse à ces questions peut être donnée *ex ante*, dans le cadre d'une évaluation explorant s'il est judicieux de généraliser le déploiement d'un programme mis en œuvre dans un périmètre géographique donné, ou *ex post*, dans le cadre d'une évaluation qui détermine si une politique publique a eu les effets économiques escomptés sur une population donnée au cours d'une période de temps définie.

L'avantage principal de cette méthode est qu'elle permet d'envisager de façon exhaustive tous les effets économiques possibles d'une politique publique sur une population donnée. Son utilisation permet donc d'améliorer la transparence sur critères de décision des politiques publiques. En utilisant les résultats d'une évaluation coût-efficacité, le décideur public peut arbitrer non seulement en fonction de critères économiques (coûts), mais également en fonction de critères d'efficacité des résultats de l'action publique. Autrement dit, le recours à l'évaluation coût-efficacité favorise une prise de décision pas uniquement guidée par des considérations de contrôle des dépenses publiques, notamment

lorsqu'elle utilise des critères d'efficacité mesurés en termes de bien-être individuel. Cette méthode est d'ailleurs identifiée par France Stratégie comme un outil central pour comparer l'efficacité de différentes politiques publiques (Desplatz et Ferracci 2016).

III. Des exemples d'utilisation de cette méthode dans l'évaluation des politiques de santé

Récemment, de nombreux travaux scientifiques se sont intéressés à l'évaluation de l'efficacité des politiques de lutte contre la pandémie de COVID-19. Une revue systématique de cette littérature identifie ainsi que les principales mesures de lutte contre le COVID-19 (tests, port du masque, distanciation sociale, quarantaines) ont été majoritairement efficaces, c'est-à-dire que leur retour sur investissement a été élevé, et ce d'autant plus que le facteur de reproduction du virus était élevé et qu'elles étaient introduites en combinaison (Vandepitte et al. 2021). Cette étude alerte néanmoins sur l'existence de facteurs propres aux pays (densité et structure de la population, organisation du système de santé, etc.) qui expliquent une efficacité plus forte de ces mesures selon les pays. Elle montre que les résultats d'une analyse d'efficacité d'une politique réalisée dans un pays/contexte particulier ne peuvent pas être aisément transposés à un autre pays.

L'évaluation coût-efficacité peut aussi être utilisée de façon prospective pour guider la décision publique. Par exemple, depuis plus de dix ans, la Haute Autorité de Santé (HAS) utilise cette méthode d'évaluation pour déterminer l'efficacité des innovations de santé et éclairer les négociations de fixation de prix et de remboursement de ces innovations entre les industriels de santé et le Comité économique des produits de santé (CEPS). Elles sont publiées de façon transparente sur le site de la HAS. Ces analyses sont l'un des principaux outils utilisés par le CEPS pour évaluer l'impact attendu d'une décision de fixation de prix d'un

médicament. Les industriels ont la responsabilité de produire les modèles d'évaluation de leurs innovations, en suivant un guide méthodologique conçu et actualisé par la commission d'évaluation économique et de santé publique (CEESP) de la HAS. Ce guide détaille précisément les critères retenus pour évaluer la qualité de l'analyse (HAS 2020). Une fois l'analyse réalisée, les industries déposent un « dossier d'efficience » qui décrit le contenu du modèle et ses résultats. Des « avis d'efficience » rendus par la CEEESP concluent sur l'impact de l'efficience liés à l'introduction d'un nouveau traitement sur le marché français, ou évaluent *ex post* quelle a été l'efficience d'un traitement après plusieurs mois d'utilisation.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

La qualité de cette méthode dépend de celle des données utilisées pour construire le modèle, et de la qualité de la méthode d'identification de la relation de causalité pour mesurer les effets de la politique. Ces deux points sont essentiels. En effet, il est primordial que l'étape d'estimation des effets de la politique mobilise des méthodes d'identification avancées (essai randomisé, scores de propension, variables instrumentales) qui sont souvent complexes à mettre en œuvre. Lorsque ces données ne sont pas disponibles, il faut mobiliser les données de la littérature scientifique pour trouver des évaluations comparables dans d'autres pays, et utiliser les données de ces évaluations pour « nourrir » le modèle. Si ces données sont absentes de la littérature, on doit avoir recours à des entretiens qualitatifs, ce qui peut réduire la précision des hypothèses et la qualité générale du modèle.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

L'avantage principal de l'analyse coût-efficacité est qu'elle représente un outil d'aide à la décision politique transparent et facile d'accès. En effet, la comparaison de l'efficacité de différents programmes est réalisée à l'aide d'une représentation graphique qui permet facilement d'identifier les mesures les plus efficaces, c'est-à-dire celles dont le ratio coût-efficacité est le plus favorable car elles sont moins chères qu'un comparateur pour une efficacité plus forte. De plus, ces méthodes d'analyses sont robustes : elles sont utilisées depuis plusieurs décennies dans tous les domaines de l'économie (santé, développement, éducation, travail, etc.).

Néanmoins, cette méthode a deux limites principales. La première limite est liée au fait que cette méthode ne permet pas réellement de guider la décision publique lorsque le résultat de l'évaluation montre que la politique est plus efficace mais aussi plus coûteuse qu'une autre mesure. C'est souvent le cas lorsque l'on évalue l'impact d'une politique visant à encourager le déploiement d'une innovation sur un marché, qui est souvent plus chère mais aussi plus efficace qu'un comparateur. La seconde limite de cette méthode est liée au coût parfois élevé de sa mise en œuvre. La mise en œuvre d'un essai randomisé nécessite un investissement financier important (une « petite » expérimentation peut coûter plusieurs centaines de milliers d'euros). De plus, l'estimation de l'effet de la politique implique un suivi temporel long, qui est souvent déconnecté du temps politique. C'est pour cette raison que la faisabilité des évaluations d'efficacité dépend souvent de la capacité de l'organisme d'évaluation à mener une expérimentation. Cela a pu par le passé bloquer la mise œuvre de mesures publiques. Par exemple, le déploiement de la télémédecine en France, bien que souhaité par les pouvoirs publics,

a longtemps été bloqué par l'incapacité des acteurs du marché et/ou de l'Assurance maladie à mener des expérimentations permettant de conclure à l'efficacité de ces dispositifs.

Quelques références bibliographiques pour aller plus loin

Desplatz, Rozenn. et Ferracci, Marc. 2016. « Comment évaluer l'impact des politiques publiques? Un guide à l'usage des décideurs et praticiens ». Paris, France: France Stratégies.

HAS. 2020. *Choix méthodologiques pour l'évaluation économique*. Saint-Denis.

Vandepitte, Sophie. et Alleman, Tijs. et Nopens, Ingmar. et Baetens, Jan. et Coenen, Samuel. et De Smedt, Delphine. 2021. « Cost-Effectiveness of COVID-19 Policy Measures: A Systematic Review ». *Value in Health* 24 (11): 1551-69. <https://doi.org/10.1016/j.jval.2021.05.013>.

PARTIE II

MÉTHODES QUALITATIVES

9. Observation directe et ethnographie

NICOLAS FISCHER

Résumé

L'observation directe ou ethnographie est une méthode qualitative consistant à observer directement sur le terrain la situation sociale que l'on cherche à étudier — par exemple la mise en œuvre d'une politique publique — ce qui implique que le chercheur ou la chercheuse soit physiquement présent·e sur le terrain. C'est une méthode exigeante sur le plan de l'engagement qu'elle demande (présence physique durable sur le terrain, prise de notes systématique). Elle permet de rendre compte de la réalité des pratiques et des interactions, à distance des discours officiels.

Mots-clés : Méthodes qualitatives, ethnographie, observation directe, mise en œuvre des politiques publiques, entretien semi-directif, interactions, étude de cas

I. En quoi consiste cette méthode?

L'observation directe s'inspire de l'observation ethnographique pratiquée de longue date en sciences sociales, notamment en anthropologie. Elle relève des méthodes d'évaluation qualitatives. Elle vise ainsi à combler les limites des enquêtes quantitatives, fondées uniquement sur des analyses statistiques : ces dernières proposent en effet un tableau chiffré global

des résultats d'une politique, mais elles ne disent rien en revanche des modalités de sa mise en œuvre, et des difficultés concrètes qui sont responsables de ses échecs ou de ses effets imprévus. L'observation directe permet au contraire de saisir directement sur le terrain les situations pratiques de mise en œuvre d'une politique publique : on dispose alors d'une description de première main de la concrétisation d'un programme de politique publique, mais aussi des conditions matérielles de son succès ou de son échec.

L'observation directe des pratiques sociales possède une histoire ancienne. Elle est tout d'abord inséparable de l'anthropologie et de l'ethnologie : lorsqu'elles se constituent pleinement en sciences au cours du XIX^{ème} siècle, ces disciplines théorisent progressivement l'ethnographie comme leur principale méthode de collecte des données. Alors qu'il s'agit à l'époque d'étudier des populations éloignées géographiquement et culturellement, l'observation permet de réduire la distance sociale avec les sujets de l'enquête à travers une recherche en immersion, qui suppose des séjours prolongés sur place, l'apprentissage des langues locales, et une série de précautions méthodologiques destinées à éviter tout jugement de valeur ethnocentrique de la part de l'ethnographe. À la fin du XIX^{ème} siècle, et dans une perspective plus proche de la démarche d'évaluation, les enquêtes sociales menées en Europe auprès des populations ouvrières ou marginalisées recourent également à l'observation, sensée là encore réduire la distance sociale séparant l'ethnographe du milieu qu'il ou elle observe. Enfin, au XX^{ème} siècle, l'observation est mobilisée en sociologie, et plus tard en science politique, pour étudier cette fois des objets « proches » (services publics, partis politiques, associations). L'enjeu est de « dé-familiariser » ces pratiques connues, la position d'observateur·rice invitant à décentrer le regard et à interroger les causes et les ressorts sociaux d'activités qui paraissent aller de soi.

Au sein de la famille qualitative, l'observation est souvent combinée avec la réalisation d'entretiens semi-directifs (voir chapitre séparé sur l'entretien semi-directif), tant avec les fonctionnaires qu'avec les publics

qu'ils et elles rencontrent. Là encore, l'observation permet de restituer ce que ces entretiens ne peuvent dire : elle permet tout d'abord de contourner l'autocensure que s'imposent souvent les personnes rencontrées en entretien, notamment lorsqu'il s'agit d'évoquer la qualité de leur travail et la réalisation de leurs missions. Elle permet également de décrire précisément certaines dimensions de l'action publique que les personnes conduisant l'évaluation et les personnes évaluées ne songeront pas à évoquer en entretien. Les routines et habitudes locales, l'organisation pratique du travail, les postures et attitudes ou la communication non-verbale avec les usagers – et tout ce qu'elles révèlent des rapports sociaux et des inégalités engagées dans la relation entre les fonctionnaires et leurs publics – se rendent alors directement visibles (Perret, 2008). Ce type d'approche peut s'avérer particulièrement utile lorsque les politiques évaluées s'adressent à des populations sensibles (personnes précaires ou socialement marginalisées, personnes handicapées...) dont la prise en charge suppose de la part des fonctionnaires des compétences interactionnelles particulières : présentation de soi, capacité à expliquer les démarches à effectuer ou à gérer l'anxiété ou la colère des publics rencontrés.

La mise en œuvre d'une enquête ethnographique nécessite une préparation particulière (Becker, 2002). S'il peut paraître aisé de se rendre dans un lieu pour l'observer, il convient en effet de produire le regard particulier de l'observateur·rice, et de constituer ainsi le·s espace·s étudié·s en scène d'observation. Un travail théorique et documentaire important est ainsi indispensable pour le repérage des sites d'observation pertinents : quel guichet, quels bureaux observer, dans quel lieu (commune rurale, urbaine, riche ou défavorisée)? Sur quelles activités, quelles dimensions se concentrer? Faut-il chercher à comparer la même phase de l'action publique dans des lieux différents, ou au contraire analyser les différentes étapes d'une chaîne administrative? Après avoir répondu à ces questions, l'ethnographe doit se rendre sur le terrain, et s'y confronter à la tension inévitable entre rapprochement et distanciation vis-à-vis des enquêté·e·s. L'observation suppose en effet de partager le

quotidien des personnes enquêtées au cours d'une longue période, en réduisant au maximum la distance qui nous sépare potentiellement d'elles. Il s'agit donc d'aligner autant que possible son apparence vestimentaire, son élocution et son *hexis* corporelle sur celle des enquêté·e·s. À l'inverse, il convient également de quitter régulièrement le terrain d'observation pour se « retrancher » dans un espace propre à la réflexion sur les activités observées : il s'agit dans ce cas d'éviter une trop forte immersion dans la pratique, et de refonder ainsi la position extérieure d'observation.

Tout au long de l'observation, les activités observées sont consignées régulièrement dans un journal de terrain, sous une forme écrite ou enregistrée. S'il n'existe aucune forme ou méthode standardisée pour sa rédaction, ce journal doit combiner non seulement la description (des lieux observés, avec plans et croquis, et des activités qui s'y déploient), mais aussi les réactions de l'ethnographe : la surprise, l'indignation ou la sympathie face aux phénomènes observés informent en effet sur la sensibilité de l'observateur·rice, mais aussi sur celle, divergente, des enquêté·e·s : elle met en évidence la production de représentations locales de ce qui est « normal », « acceptable » ou « problématique », représentations que ne partage pas (encore) une personne extérieure qui découvre la situation. D'un point de vue méthodologique, consigner ses réactions au fil de l'observation permet également de les objectiver pour les analyser, en limitant ainsi l'impact de la subjectivité de l'ethnographe sur ses observations.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Comme le notent Stéphane Beaud et Florence Weber (2012), l'adoption de la méthode ethnographique résulte d'une insatisfaction vis-à-vis du discours qu'un groupe — ici une administration — tient sur lui-même :

il s'agit d'aller au-delà de la présentation officielle d'une activité, de ce qu'en disent les règles de droit, les instructions ou les plaquettes de présentation, pour analyser la réalité de sa pratique. Une telle observation directe peut donc avoir lieu *ex post*, au stade de la mise en œuvre des politiques publiques, dont on sait qu'elle correspond souvent à une véritable ré-élaboration de l'action publique par les fonctionnaires de terrain. Elle se justifie particulièrement lorsqu'il s'agit d'évaluer un format d'action publique difficilement quantifiable (l'accueil au guichet d'une administration par exemple, cf. section suivante). Une telle démarche permet alors d'observer la diversité des investissements locaux d'une même politique, et son adaptation aux conditions locales de sa mise en œuvre (spécificité des publics, du contexte socio-économique ou politique) ou des acteurs qui l'assurent (héritage des routines locales propres à une direction, à un bureau ou à une commune). Une telle perspective ouvre sur deux logiques évaluatives potentielles : mettre en évidence les innovations locales dont sont capables les fonctionnaires pour traiter des situations non prévues par les textes, et envisager également les multiples logiques qui peuvent éventuellement faire dévier une politique publique de son objectif affiché. Il s'agit alors typiquement d'évaluer l'ajustement d'une politique et des moyens matériels qui lui ont été alloués avec les réalités rencontrées sur le terrain, de repérer les enjeux négligés lors de sa conception, et d'isoler les pratiques à modifier pour permettre à l'action publique de produire pleinement ses effets.

III. Un exemple d'utilisation de cette méthode : l'évaluation de la politique d'accueil dans les services publics

Bien qu'il soit déjà ancien, le rapport remis en 1993 au Premier Ministre sur *Les services publics et les populations défavorisées : évaluation de la politique d'accueil* (Paris: la Documentation française, 1993) constitue un bon exemple de mise en œuvre de la méthode ethnographique pour

l'évaluation. Il illustre tout d'abord l'intérêt de l'observation pour opérer une approche fine de la question initialement posée, en 1990, par le Comité interministériel de l'évaluation : dans un contexte marqué par l'essor du thème de la modernisation des services publics et par la mise en place de la Politique de la ville, l'enjeu était d'évaluer la capacité des guichets locaux des services publics à se saisir effectivement des difficultés rencontrées au quotidien par les populations les plus précaires. Une telle analyse ne pouvait passer ni par une évaluation purement quantitative, ni par une simple enquête par entretien : l'objectif était bien en effet de s'intéresser à des interactions – celle des services de l'État situés en « première ligne » avec leurs publics les plus dépendants des prestations qu'ils assurent – et de tenter d'évaluer leur qualité – notamment pour juger de la capacité des usagères et usagers à faire valoir effectivement leurs droits. Il s'agissait ainsi de s'interroger sur la mise en place de l'accueil, la qualité de l'information des publics, l'impact sur l'effectivité de leurs droits, la possibilité de mettre en place des indicateurs de satisfaction et *in fine*, sur l'opportunité d'adopter des politiques sélectives en matière d'accueil, dont certaines seraient adaptées aux publics défavorisés.

Ce rapport met également en évidence la combinaison le plus souvent nécessaire de l'observation avec d'autres méthodes permettant d'éclairer les constats ethnographiques et de monter en généralité : en l'occurrence, l'enquête qualitative est combinée avec un volet quantitatif (le passage de questionnaires auprès des usagers ayant permis des tris croisés portant sur leurs caractéristiques socio-démographiques). Au sein même du volet qualitatif, les observations menées au guichet sont complétées par la réalisation d'entretiens qualitatifs avec des usagers, des agents d'accueil et des « intermédiaires sociaux » (acteurs associatifs ou fonctionnaires des services sociaux facilitant l'accès aux services publics).

La mise en place de l'enquête a donc supposé le travail conjoint des services d'inspection de l'administration et de bureaux d'étude ou centres de recherche (3 bureaux privés et un centre universitaire), et un travail préalable de repérage des scènes d'observation pertinentes : chaque

enquête est préparée par un travail de cartographie de l'ensemble des services urbains, qui permet de dégager huit services publics jugés centraux dans la problématique de l'accueil (police, urgences des hôpitaux, mairie...). Les localités enquêtées ont été sélectionnées en raison de leur classement préexistant comme « zones défavorisées ».

Ces choix méthodologiques ne sont pas dépourvus de biais, et illustrent au passage une des difficultés de l'enquête ethnographique et l'importance conjointe de la question initialement posée, et du protocole d'observation mis en œuvre pour y répondre. En l'occurrence, le rapport conclut à la nécessité d'adapter les politiques d'accueil aux populations défavorisées, notamment par la création de plates-formes ou « maisons des services publics » réunissant dans un même lieu, au sein de quartiers marginalisés, des permanences de différents services publics (Poste, mairie, etc.). Ces conclusions ont été critiquées par des universitaires ayant mené leurs propres enquêtes ethnographiques sur les usagers précaires des guichets (voir Siblot, 2005; également Dubois, 2003) : en mettant l'accent sur la seule dépendance des usagers vis-à-vis des services publics, l'évaluation reste aveugle selon eux aux multiples stratégies de « débrouille » que les populations précaires sont à même de développer pour faire valoir leurs droits, et que révèle une enquête ethnographique approfondie. De même, l'évaluation effectuerait une généralisation abusive en affirmant le caractère dominé des usagers, alors qu'ils sont inégalement dotés en capitaux, scolaires notamment, et peuvent être à même pour certains d'entre eux d'interagir sur un pied d'égalité avec les agents d'accueil.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

L'observation ethnographique sera d'autant plus utile que les observateur·rice·s auront été capables de faire un travail de *casing* : c'est-à-dire de constituer les situations toujours singulières observées sur le terrain en « cas » pouvant corroborer ou invalider une théorie, et permettant donc de traiter une problématique spécifique. L'enjeu est alors de « borner empiriquement une relation problématique entre des idées et des preuves, entre la théorie et les données » (Hamidi, 2012). Maintenir cette relation suppose des ethnographes une attention constante aux pratiques observées sur le terrain : elles font régulièrement surgir des logiques ou des thématiques imprévues, qui doivent conduire à enrichir ou modifier parfois notablement la problématique théorique de départ. L'enjeu est important dans la perspective d'une évaluation de l'action publique, où ce sont les attendus de départ de la mission d'évaluation qui peuvent alors être modifiés, sous peine de négliger certaines réalités du terrain (problème notamment soulevé au sujet de l'évaluation de l'accueil des populations défavorisées dans les services publics, cf. section précédente).

La complexité de l'exercice ethnographique réside alors dans la capacité des observateurs et observatrices à articuler, dans une même recherche, des cas de statuts différents (Hamidi, 2012, en référence à l'*extended case theory* de l'école de Manchester). On peut ainsi associer des cas « exemplaires » pour lesquels on peut s'attendre, compte tenu du contexte et des populations concernées, à ce que les hypothèses théoriques soient pleinement validées (pour conserver l'exemple précédent : un guichet de la Poste dans un quartier populaire d'une zone urbaine délaissée), et des cas « limites » dans lesquelles elles ne seront que partiellement confirmées (un autre guichet situé dans un quartier moins enclavé, ou situé dans une zone aux solidarités communautaires ou au tissu associatif plus étroits). Les différents facteurs qui peuvent

influencer la mise en œuvre d'une politique sont inégalement présents dans ces différents cas : leur rapprochement permet dès lors d'identifier avec finesse ceux qui jouent pleinement sur l'action publique, et ceux qui sont plus secondaires.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

On l'a vu, l'observation directe permet particulièrement de saisir *ex post* les conditions matérielles de la mise en œuvre d'une politique sur le terrain, en se détachant des présentations officielles. Le repérage de scènes d'observation présentant chaque fois des conditions différentes de mise en œuvre des politiques publiques, peut permettre une évaluation particulièrement fine des effets d'une politique donnée.

On l'a vu également, l'observation a le plus souvent vocation à être croisée avec d'autres méthodes et des approches complémentaires. Un reproche classique adressé à l'observation directe concerne en effet la possibilité de généraliser ses résultats : les observations, effectuées dans un espace déterminé et nécessairement situées, ne concerneraient que le contexte local qu'elles décrivent et ne permettraient pas de passer de l'échelle microsociologique à l'échelle macro, celle d'une évaluation plus globale de la politique publique concernée. Cette objection est en partie dépassée dans les travaux récents, qui ont souligné la nécessité de compléter l'enquête ethnographique par une analyse mobilisant d'autres méthodes, afin de relier les pratiques observées localement avec leur cadre institutionnel et son histoire. Ce lien peut s'établir différemment en fonction des approches : dans l'enquête menée par Vincent Dubois (2003) sur les guichets des caisses d'allocations familiales (CAF), c'est la réalisation d'entretiens avec les guichetières et guichetiers qui permet de relier l'observation des interactions au guichet avec le parcours professionnel des fonctionnaires, et au-delà avec les conditions

institutionnelles de leur recrutement (absence de définition claire de la mission des guichetier-e-s et de fiche de poste, etc.). Sur la même thématique, l'enquête de Jean-Marc Weller (1999) met l'accent sur l'organisation matérielle de l'accueil dans les administrations et ce qu'elle révèle (coupes budgétaires, retrait de l'État providence et nouvelle conception managériale qui fait des usagèr-e-s des « client-e-s ») pour relier les interactions observées sur le terrain à des réformes globales de l'action publique, dont elles sont le reflet.

Une autre limite de la méthode ethnographique est l'investissement en temps et en personnel qu'elle demande. Si l'observation est peu coûteuse techniquement — elle ne requiert ni matériel d'enregistrement, ni traitement informatique des données collectées — elle suppose en revanche la présence d'un-e observateur-ric-e, ou plus souvent d'un groupe d'observateur-ric-e-s agissant de manière concertée sur plusieurs scènes et pour de longues séquences d'observation (plusieurs mois), en alternant les périodes de « retrait » puis de « retour » sur le terrain. Il s'agit alors à la fois de saisir les évolutions des pratiques (notamment lorsqu'il s'agit d'évaluer la mise en œuvre d'une réforme récente, que les fonctionnaires de terrain découvrent puis se réapproprient progressivement), mais aussi, on l'a vu, d'autoriser les évaluateur-ric-e-s à se soustraire régulièrement au travail de terrain pour confronter leurs conclusions dans le cours même de l'enquête, et préciser ou modifier le constat général qu'ils et elles entendent faire sur la politique évaluée. Si ce temps long de l'enquête peut dès lors paraître chronophage, on voit qu'il ne renvoie pas seulement au travail de « terrain » et à l'observation : il correspond aussi à un temps de (ré)élaboration du rapport d'évaluation final, et des conclusions générales qu'il proposera.

Quelques références bibliographiques pour aller plus loin

Beaud, Stéphane. 2010. *Guide de l'enquête de terrain : produire et analyser des données ethnographiques*, Grands Repères Guides. Paris: La Découverte.

Dubois, Vincent. 2003. *La vie au guichet. Relation administrative et traitement de la misère*. Paris: Economica.

Hamidi, Camille. 2012. « De quoi un cas est-il le cas? Penser les cas limites ». *Politix*, n°100, vol. 4: 85-98.

Jeannot, Gilles. 2008. « Les fonctionnaires travaillent-ils de plus en plus? Un double inventaire des recherches sur l'activité des agents publics ». *Revue française de science politique* 58, n°1: 123-40.

Siblot, Yasmine. 2005. « “Adapter” les services publics aux habitants des “quartiers difficiles”. Diagnostics misérabilistes et réformes libérales ». *Actes de la recherche en sciences sociales*, 159, n°4: 70-87.

Weller, Jean-Marc. 1999. *L'État au guichet. Sociologie cognitive du travail et modernisation administrative des services publics*. Paris: Desclée de Brouwer.

10. L'entretien semi-directif

CLÉMENT PIN

Résumé

Technique d'enquête qualitative très répandue, l'entretien semi-directif consiste en une interaction verbale sollicitée par l'enquêteur·rice auprès d'un·e enquêté·e, à partir d'une grille de questions utilisée de façon très souple. L'entretien vise à la fois à collecter des informations et à rendre compte de l'expérience de la personne et de sa vision du monde, dans une optique compréhensive. Il est utile pour différents types d'évaluations des politiques publiques, et notamment pour clarifier les objectifs d'une politique, analyser sa mise en œuvre ou encore étudier sa réception.

Mots-clés : Méthode qualitative, entretien semi-directif, induction, empathie, étude de cas, idéaux-types, évaluation réaliste

I. En quoi consiste cette méthode?

L'entretien semi-directif est une technique de recueil de données très largement utilisée dans la recherche qualitative en sciences sociales. De manière très générale, elle se distingue radicalement de l'enquête par questionnaire qui vise à produire des données standardisées sur une vaste population pour rechercher par traitement statistique des régularités dans la variation des opinions ou des attitudes entre groupes d'individus. La pratique de l'entretien, quelle que soit sa forme spécifique, sert quant à elle à produire des données permettant avant tout de mieux saisir la singularité de l'expérience que des individus ou groupes d'individus ont de leurs relations avec les autres, avec les institutions, ou plus largement

celle qu'ils ont de phénomènes sociaux. Si l'étude qualitative et approfondie du singulier peut donner lieu en elle-même à des connaissances ayant un certain degré de généralisation, celles-ci procèdent d'un traitement de données raisonnant par étude de cas et idéaux-types, ainsi que par recoupement avec des données recueillies au moyen des deux autres techniques qualitatives classiques que sont l'observation et le traitement de sources écrites. Les techniques qualitatives peuvent également être mobilisées dans le cadre de recherches adoptant une méthodologie mixte.

La pratique de l'entretien a émergé au XIX^e siècle dans le champ de la psychologie clinique et de l'enquête sociale à des fins respectivement médicales et politiques. Elle s'est développée en tant que technique de recherche à part entière au cours du XX^e siècle aux États-Unis puis en Europe dans une démarche de sociologie compréhensive dans la filiation des travaux de Max Weber. La fonction de l'entretien est ainsi de recueillir la parole des individus, le postulat théorique général étant que les phénomènes sociaux ne peuvent être compris et donc expliqués indépendamment du sens que les individus donnent à leurs actions. Sur cette base commune, plusieurs pratiques scientifiques de l'entretien ont été progressivement formalisées, les principales étant l'entretien ethnographique, l'entretien non-directif, et l'entretien semi-directif. C'est toutefois ce dernier qui s'est imposé au cours des dernières décennies, en particulier en France, comme la technique la plus utilisée en sociologie de l'action publique (Pinson, Sala Pala, 2017). Dans ce cadre, il est souvent employé si ce n'est comme un mode exclusif de recueil de données, du moins comme un mode privilégié, au motif qu'il permet de produire des données ayant une valeur intrinsèque (et pas seulement par recoupement avec des observations ou de la documentation).

Comme les autres formes d'entretien en sciences sociales, l'entretien semi-directif est une interaction verbale sollicitée par l'enquêteur-riche auprès d'un-e enquêté-e. Mais, dans le cas de l'entretien semi-directif, la situation d'interaction a ceci de particulier que l'enquêté-e est de prime

abord placé·e dans un rôle d'informateur·rice, de détenteur·rice d'un savoir (commun, non scientifique) précieux sur le thème d'intérêt de l'enquêteur·rice.

Sur le plan épistémologique, l'entretien semi-directif s'inscrit dans un mode de raisonnement scientifique où le terrain (donc ce que disent les enquêté·e·s) n'est pas qu'une instance de vérification de théories élaborées dans l'abstrait, mais bien ce à partir de quoi s'engage l'élaboration de la question de recherche et des hypothèses : la théorie est produite par induction à partir des données de terrain, selon le principe de la *Grounded theory* popularisé par Anselm Strauss.

Qu'entend-on par entretien « semi-directif » ? S'il convient pour l'enquêteur·rice de préparer une grille organisée de questions qui lui servira de guide pour orienter l'entretien, l'usage de cette grille n'est pas rigide. L'enjeu est que l'enquêté·e fournisse par ses prises de parole le plus d'informations tant objectives (sur les phénomènes, institutions ou processus étudiés) que subjectives (sur ses représentations, son système de valeurs, ses croyances). Il convient dès lors d'interagir avec la personne interviewée de sorte que celle-ci en vienne à endosser activement son rôle d'informatrice, dans une logique de conversation plutôt que de questionnaire administré « de haut ». La qualité d'un entretien semi-directif dépend ainsi en grande partie de l'attitude d'empathie et d'écoute attentive adoptée par l'enquêteur·rice, qui lui permettra de faire l'usage le plus adapté de sa grille de questions en situation (Kaufmann, 2016).

L'application de ces principes méthodologiques n'aura jamais pour effet de couper court aux débats propres au champ de la recherche qualitative et sur les différentes formes d'entretiens, que ces débats portent sur la validité des données recueillies (leur degré d'objectivité/subjectivité, leur véracité/facticité, leur partialité etc.), ou entre paradigmes scientifiques (constructivisme/réalisme critique), si bien qu'il n'y a de bon usage de l'entretien semi-directif qui ne soit réfléchi, méthodiquement élaboré, et explicite.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

La réalisation d'entretiens semi-directifs peut servir à traiter trois grands types de questions évaluatives. Elle peut en premier lieu aider à rendre intelligible l'ensemble souvent complexe des objectifs initiaux d'une politique publique. Les entretiens semi-directifs peuvent ensuite être utilisés dans une démarche d'évaluation visant à retracer les processus de mise en œuvre de la politique, de comprendre comment ses objectifs se traduisent concrètement dans des interventions et des pratiques des agents administratifs. Enfin, bien que moins reconnue pour cela dans le contexte français, l'enquête par entretiens semi-directifs peut contribuer à produire des évaluations en documentant la réception d'une politique par ses bénéficiaires et plus largement par ses destinataires. Si ces trois usages peuvent être combinés dans une même recherche évaluative, nous précisons successivement leurs apports respectifs.

Dans la perspective de clarification des objectifs d'une politique, l'entretien semi-directif apparaît comme un des rares moyens d'approcher empiriquement le travail gouvernemental et plus précisément les processus de décision intervenant dans la mise à l'agenda de problèmes publics et la définition de programmes d'action pour les traiter. En raison de leur caractère hautement politique, les sphères gouvernementales restent d'un accès difficile pour conduire des observations. Les sources écrites, par leur caractère officiel et consensuel, restent quant à elles pauvres en informations pour saisir les débats et controverses entre acteurs décisionnels mus par des idéologies, des logiques institutionnelles et des intérêts particuliers. L'entretien semi-directif est alors utilisé comme technique permettant d'accéder de manière rétrospective à des informations de première main indispensables pour décoder les enjeux ayant présidé à la formation des compromis et arbitrages ne s'exprimant que très implicitement dans la formulation officielle des objectifs d'une politique.

Dans une démarche d'évaluation centrée sur l'étude des moyens effectivement déployés (*outputs*) en application d'une politique, l'usage de l'entretien semi-directif apparaît au premier abord moins central. D'une part, les données nécessaires ayant par définition un caractère administratif et technique prononcé, elles sont souvent disponibles sous forme écrite. De plus, les pratiques des agentes et des agents étant considérées comme plus ordinaires, elles se prêtent davantage à l'observation. L'enquête se fait alors plus ethnographique, de manière à saisir les pratiques d'adaptation de la règle à la diversité des situations et des publics concernés (voir chapitre séparé sur l'observation directe). L'entretien semi-directif peut toutefois être mobilisé en complément pour croiser les hypothèses explicatives portant sur les pratiques des agent-e-s avec le récit qu'ils et elles font de leurs situations de travail et les représentations expertes qu'ils et elles élaborent au sujet de « leurs » publics.

L'usage de l'entretien semi-directif dans l'étude des effets (*outcomes*) d'une politique se conçoit dès lors qu'on ne réduit pas cette étude à la seule mesure (quantitative) des impacts mais que l'on cherche à comprendre (qualitativement) le processus de production de ces effets. Ce type d'analyse, formalisé dans les années 1990 par les pionniers de l'évaluation qualitative tels que Michael Patton, souligne qu'une même politique peut revêtir des significations différentes selon les populations concernées, et que cette diversité produit une variation importante dans ses effets. Le concept de réception (Revillard, 2019) aide à analyser les interactions entre les logiques d'appropriation (cognitives et pratiques) et les effets (symboliques et matériels) d'une politique. L'étude empirique de la réception passe par la réalisation d'entretiens semi-directifs dont la particularité est d'accorder la primauté à la dimension compréhensive plutôt qu'informative, l'examen portant en premier lieu sur la subjectivité des destinataires. Une autre pratique de l'entretien semi-directif, moins subjectiviste, est également développée dans la démarche d'évaluation propre au courant de l'évaluation réaliste. Nous la présentons dans la section suivante.

III. Un exemple d'utilisation de cette méthode en évaluation des politiques éducatives

Théorisée par le sociologue Ray Pawson, la démarche d'évaluation réaliste est aujourd'hui bien reconnue dans la littérature scientifique internationale et est mobilisée par de nombreuses organisations gouvernementales (voir chapitre séparé sur l'évaluation réaliste). Sa caractéristique première est de substituer à la question ordinaire « cette politique fonctionne-t-elle? » (au sens de produit-elle les effets recherchés?) un questionnement plus circonstancié sur « quels effets produit-elle? pour qui? dans quels contextes? À quelles conditions? ». Le réalisme (critique) de cette démarche réside dans le postulat que la mesure d'impact d'une politique est insuffisante pour en saisir ses effets, que ceux-ci sont tellement différents selon ses destinataires et le contexte qu'il est indispensable, pour l'évaluer, de comprendre la variété des processus qu'elle active. Évaluer une politique consiste dès lors à formuler et à examiner empiriquement des hypothèses sur la manière dont interagissent les contextes, les mécanismes et les effets (schéma d'analyse CMO, *contexts-mechanisms-outcomes*).

Le travail de formulation et d'examen des hypothèses s'appuie de manière centrale sur la réalisation d'entretiens semi-directifs conçus selon une logique qualifiée de *teacher-learner function* (Pawson, 1996), à mi-chemin entre l'entretien directif (*structured*) et non directif (*unstructured*). La dimension informative de l'entretien est dominante, l'échange avec l'interviewé-e consistant moins à partir de son vécu et de ses représentations qu'à alimenter une réflexion sur des hypothèses de recherche (*theory-driven*). Cette pratique d'entretien ne peut toutefois pas être qualifiée de directive dans la mesure où selon les phases de l'enquête, l'enquêteur-riche et l'enquêté-e vont jouer alternativement les rôles d'enseignant-e et d'apprenant-e. Pour aider à anticiper et maîtriser cette permutation des rôles, Ana Manzano (2016) distingue trois phases de l'enquête par entretiens. Le premier ensemble d'entretiens remplit

une fonction de glanage de théories (*theory gleaning*), c'est-à-dire de recensement auprès des acteur·rices d'hypothèses provisoires sur les effets des circonstances contextuelles sur le fonctionnement du programme étudié. Dans une deuxième phase, certaines théories ont été écartées, et les théories sélectionnées sont examinées plus en détail au moyen d'entretiens se faisant moins standardisés pour interroger diversement les interlocuteur·rices dans une visée d'affinage théorique (*theory refining*). C'est surtout dans la troisième phase de consolidation théorique (*theory consolidation*) que l'évaluateur·rice se fait *enseignant·e* en exposant à l'enquêté·e sa compréhension contextualisée du programme, auquel l'enquêté·e peut réagir en mobilisant des exemples dans une logique de vérification ou falsification.

Un exemple récent d'évaluation conduite dans le champ des politiques éducatives illustre particulièrement bien cette pratique de l'enquête par entretien semi-directif. Afin d'évaluer la politique colombienne visant à réduire les inégalités régionales de réussite scolaire en étendant de manière universelle la durée des journées d'école (programme *Jordana Unica*), Juan David Parra (2022) a réalisé une enquête qualitative comportant 31 entretiens (11 avec des responsables de services centraux et déconcentrés de l'État, 20 avec des directeurs et des éducateurs dans les écoles), 20 *focus groups* (10 avec des parents, 10 avec des élèves) et 40 heures d'observations non participantes dans des écoles. Il a également administré un questionnaire auprès d'un échantillon représentatif de directeurs d'école (N = 681). Cette enquête lui a permis de formuler, affiner puis consolider des hypothèses sur la mise en œuvre, la réception et les effets de cette politique en soulignant l'importance de raisonner à trois niveaux : celui des logiques de décentralisation des politiques éducatives, du bien-être des enfants et des adolescents, et de la motivation des élèves.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Un premier élément conditionnant la qualité d'une enquête par entretiens semi-directifs concerne le nombre et le choix des personnes interviewées. La représentativité de l'échantillon n'étant pas un critère de validité, le principe est davantage de réaliser un nombre suffisant d'entretiens (généralement estimé entre 20 et 30) pour recueillir le témoignage de personnes qui, d'un point de vue formel ou informel, occupent des positions et se trouvent dans des situations différentes au regard de l'objet étudié, si bien qu'elles pourront avoir des points de vue différents, autrement dit des expériences, des pratiques et des représentations variées à son sujet.

Un deuxième critère de qualité se joue dans la conduite même des entretiens. L'entretien semi-directif doit permettre d'alterner des moments destinés à recueillir des narrations ou récits produits librement par l'enquêté-e (généralement au moins en début d'entretien) et des moments de plus grande directivité visant à recueillir des informations préalablement ciblées par l'enquêteur-riche. Cet art de l'entretien se prépare en amont via l'élaboration d'un guide d'entretien, évolutif au fil de l'enquête et ajustable en fonction des interlocuteurs. Ce guide ne comporte pas seulement la formulation de consignes initiales et de thèmes généraux de discussion, il établit aussi une série de relances qui permettent d'obtenir les informations recherchées. La conduite des entretiens dépend en outre de la posture qu'enquêteur-riche et enquêté-e adoptent en situation et des techniques de relance utilisées par l'enquêteur-riche.

Un troisième ensemble d'enjeux réside enfin dans le traitement des données recueillies par les entretiens. Cette étape décisive vise à analyser le contenu des entretiens de manière croisée et comparative de manière à non seulement synthétiser et vérifier par recoupement les informations recueillies, mais aussi à produire une interprétation tout à la fois globale

et circonscrite de l'objet étudié, en référence au cadre théorique et aux hypothèses de recherche initialement formulées. Cette phase de travail nécessite à ce titre de décontextualiser relativement les données recueillies au sein de chacun des entretiens en analysant leur contenu au regard de catégories d'analyse portant sur le fonctionnement du système d'action et/ou des processus étudiés et l'expérience qu'en ont les différent·e·s acteur·rice·s concerné·e·s.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

L'atout principal des entretiens semi-directifs est de fournir des données indispensables à la compréhension des processus par lesquels une politique publique produit ses effets, depuis la genèse de la multiplicité de ses objectifs et de son contenu (moyens consacrés, instruments élaborés), à ses modalités effectives de mise en œuvre et jusqu'à ses logiques variées de réception. Ces données portent sur les pratiques et les représentations de l'ensemble des acteurs et actrices impliqué·e·s ou plus largement concerné·e·s (*a priori*) par une même politique. Selon les étapes de l'enquête et les types d'enquêté·e·s sollicité·e·s (acteur·rice·s de la décision, acteur·rice·s de la mise en œuvre, bénéficiaires, destinataires) l'usage de l'entretien semi-directif peut être modulé pour activer en premier lieu sa dimension informative ou compréhensive.

Ses principales limites sont de deux ordres. Premièrement, dans le cadre d'une évaluation strictement qualitative, il est requis que l'administration de la preuve opère en croisant l'usage de l'entretien et d'autres techniques de recueil de données, à savoir l'observation et l'étude des sources écrites. Deuxièmement, en tant que méthode qualitative, il est évident que le recours à l'entretien semi-directif ne permet pas en lui-même de produire

des évaluations quantitatives, évaluations par ailleurs très utiles pour fournir des données de cadrage servant à concevoir en amont le questionnement propre à une évaluation qualitative.

Notons enfin que dans le contexte actuel de développement de l'évaluation par mesure d'impact l'enquête par entretien semi-directif peut tout à fait trouver sa place dans le cadre de recherches adoptant une méthodologie mixte (Pin, Barone, 2021). Des entretiens semi-directifs peuvent ainsi contribuer à la conception (en amont) puis à l'interprétation (en aval) d'une expérimentation randomisée. Dans ce cas comme dans d'autres, l'usage de l'entretien sera modulé selon les étapes de la recherche. La technique d'entretien semi-directif sera dans un premier temps utilisée dans une logique de « qualitatif instrumentalisé » pour aider à identifier les conditions contextuelles variées de mise en œuvre d'un programme dont on cherche à mesurer l'impact et affiner de la sorte ses modalités de mise en œuvre. L'entretien semi-directif pourra ensuite être utilisé dans une logique de « qualitatif autonomisé » pour construire des idéaux-types fournissant a posteriori des éléments explicatifs d'ordre qualitatif pour comprendre les processus causaux ayant abouti aux impacts mesurés.

Quelques références bibliographiques pour aller plus loin

Kaufmann, Jean-Claude. 2016. L'entretien compréhensif. Armand Colin.

Manzano, Ana. 2016. « The craft of interviewing in realist evaluation ». *Evaluation*, n°22: 342-360.

Parra, Juan David. 2022. « Decentralisation and school-based management in Colombia: An exploration (using systems thinking) of the Full-Day Schooling programme ». *International Journal of Educational Development*, n°91: 102579.

- Pawson, Ray. 1996. « Theorizing the interview ». *British Journal of Sociology*, n°47: 295-314.
- Pin, Clément. et Barone, Carlo. 2021. « L'apport des méthodes mixtes à l'évaluation ». *Revue française de science politique*, n°71: 391-412.
- Pinson, Gilles. et Sala Pala, Valérie. 2007. « Peut-on vraiment se passer de l'entretien en sociologie de l'action publique? ». *Revue française de science politique*, n°57: 555-597.
- Revillard, Anne. 2018. « Saisir les conséquences d'une politique à partir de ses ressortissants : la réception de l'action publique ». *Revue française de science politique*, n°68: 469-492.

II. Les focus groups

ANA MANZANO

Résumé

Les *focus groups* sont une méthode qualitative qui consiste en une conversation collective organisée avec un groupe de (généralement 4 à 8) personnes, guidée par des questions à commenter. Les *focus groups* peuvent être un moyen d'intégrer les points de vue et les expériences des utilisatrices et utilisateurs et de diverses parties prenantes dans l'évaluation d'une intervention donnée. Ils sont adaptés à différentes étapes du processus de politique publique et à différentes approches d'évaluation, souvent en combinaison avec d'autres méthodes qualitatives et/ou quantitatives.

Mots-clés : Méthodes qualitatives, *focus groups*, entretiens de groupe, évaluation participative

I. En quoi consiste cette méthode?

Les *focus groups* consistent en une ou plusieurs conversations avec un groupe de personnes rassemblées pour la discussion. Les *focus groups* sont dirigés par une chercheuse ou un chercheur (et comprennent souvent un observateur ou une observatrice), dans le but d'acquérir des connaissances sur différents résultats possibles d'interventions dans des domaines tels que la vente (*marketing*), l'influence des décisions (en matière d'orientations politiques ou de comportements de santé par exemple) ou dans le cadre de démarches de suivi et d'évaluation des politiques publiques. Des chercheurs et chercheuses formé-e-s animent

les discussions à l'aide d'une série de questions non structurées et/ou structurées que le groupe doit commenter. Ces conversations peuvent être stimulées par la mise en discussion de supports spécifiques tels que des photos, vidéos, vignettes. Des jeux peuvent également être utilisés, ainsi que des techniques de prise de décision telles que des méthodes de vote informelles. Les *focus groups* peuvent être menés sous différents formats (présentiel, virtuel, synchrone ou asynchrone en ligne) et avec des participant·e·s présentant des caractéristiques d'intérêt similaires (appelés « *focus groups* homogènes », par exemple un *focus group* composé d'enseignant·e·s) ou diverses (appelés « *focus groups* hétérogènes », par exemple un *focus group* composé d'enseignant·e·s, d'élèves et de parents).

Les *focus groups* sont une méthode de production de données qualitatives, appartenant à la famille des méthodes de discussion en groupe. Comme pour d'autres données qualitatives, il n'y a pas d'accord dans la littérature scientifique sur les méthodes de recherche quant au nombre optimal de participant·e·s à un *focus group* ou au nombre approprié de groupes à organiser. Au stade de la conception, il est utile de présenter la taille des *focus groups* sous forme de fourchettes, car de nombreux facteurs peuvent influencer sur le nombre de participant·e·s aux groupes. Certain·e·s autrices et auteurs privilégient les groupes plus petits ($n=3-5$) parce qu'ils offrent un meilleur potentiel pour explorer en profondeur des sujets complexes. Par exemple, des informations plus riches peuvent être obtenues en menant deux groupes de quatre participant·e·s qu'un groupe de huit participant·e·s. D'autres recommandent des groupes de taille moyenne ($n=6-8$), tandis que d'autres encore suggèrent des groupes plus importants ($n=6-12$) afin de capter une plus grande variété d'opinions. La durée de la discussion dépend de la taille du groupe et du sujet mais, en règle générale, 90 minutes sont nécessaires pour que tou·te·s les participant·e·s aient la possibilité d'exprimer leur point de vue. Des durées supérieures à deux heures augmentent la charge de travail des participant·e·s et risquent de les dissuader de participer.

De nombreuses expressions sont utilisées en évaluation pour décrire les méthodes de production de données à partir de groupes, telles que « focus group » et « entretien collectif » (voir fiche séparée sur les entretiens collectifs), avec des variations géographiques. Une distinction essentielle est que les *focus groups* insistent sur le rôle significatif des processus dialectiques de groupe (par exemple, les normes, la dynamique, la communication non verbale) qui peuvent aider les évaluateurs et évaluatrices à acquérir des connaissances sur les points de vue du groupe et les accords et désaccords entre sous-groupes. Les différences conceptuelles entre de nombreux termes de production de données qualitatives de groupe sont souvent peu claires et il n'y a pas toujours de consensus sur la façon dont ils diffèrent les uns des autres. Les sessions d'échange avec les parties prenantes ne suivent pas toujours le format standard du *focus group*: réunions de groupes communautaires, groupes consultatifs, événements de consultation, ateliers d'engagement public, cafés, réunions d'experts, conversations collectives facilitées avec des groupes, etc. Bien que ces sessions puissent parfois s'adresser à des groupes particuliers de bénéficiaires des interventions, elles nécessitent moins de travail préparatoire, n'impliquent pas de facilitation structurée de la part de l'équipe d'évaluation, et ne font généralement pas l'objet d'une transcription et d'une analyse formelle de contenu.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Bien que l'évaluation des politiques publiques ait été dominée par des approches expérimentales et quantitatives, les responsables politiques valorisent aussi l'implication des bénéficiaires, et les *focus groups* ont le potentiel de soutenir cet objectif participatif. Avec les entretiens approfondis, les *focus groups* sont l'une des méthodes qualitatives les plus utilisées en évaluation des politiques publiques. Plusieurs caractéristiques distinctives des *focus groups* suscitent l'intérêt des responsables

politiques, notamment le fait qu'ils permettent d'explorer les significations, les valeurs, les expériences, les points de vue et les comportements contrastés de différents sous-groupes de parties prenantes, et de saisir la complexité des contextes et des processus de mise en œuvre des politiques publiques. Les *focus groups* sont aussi souvent perçus comme ayant un intérêt politique en soi, au-delà des informations spécifiques sur les valeurs et les points de vue multiples qu'ils peuvent fournir.

Seuls ou combinés à d'autres méthodes de recherche, les *focus groups* sont utilisés dans de nombreuses approches d'évaluation (par exemple, les évaluations basées sur la théorie, les évaluations de processus et d'impact, les évaluations évolutives, participatives et celles visant le développement du pouvoir d'agir), à des fins diverses et à différentes étapes du processus politique (planification, mise en œuvre, suivi, évaluation, cycles de programmation successifs). Ils conviennent aux approches d'évaluation *ex ante* et *ex post*, et sont souvent utilisés dans les études d'évaluabilité, l'évaluations des besoins, le développement de la théorie du programme, le développement d'instruments et d'enquêtes, l'étude de la mise en œuvre, les évaluations axées sur l'utilisation et les évaluations formatives.

Les *focus groups* sont surtout utiles pour répondre aux questions d'évaluation exploratoires (pourquoi et comment) car ils offrent un moyen dynamique de décrire les politiques publiques en action. Ils permettent notamment :

- d'améliorer la compréhension d'un problème et la façon dont une intervention peut y répondre: comment le problème est perçu et vécu par les différentes parties prenantes (utilisateurs et utilisatrices, personnel de première ligne, direction), leurs attentes et les solutions qu'ils et elles proposent.

- d'obtenir un retour d'information sur la qualité, l'utilisation et la satisfaction liées aux activités et aux ressources fournies par l'intervention : ce qui a bien fonctionné, pour qui, et ce qui n'a pas fonctionné comme prévu, pourquoi et dans quelles circonstances cela s'est produit.
- de mieux comprendre le processus de mise en œuvre de la politique (par exemple, la gestion, les partenariats avec d'autres institutions/départements, la réalisation des activités et des ressources de la politique).
- de discerner les types de changements supposés/attendus (théorie du changement) et produits (le cas échéant) du point de vue des différent-e-s utilisateurs et utilisatrices et dans différents contextes politiques dans le temps et l'espace.
- d'explorer les indicateurs/critères d'évaluation lorsqu'ils ne sont pas clairs ou lorsque des critères alternatifs sont recherchés.
- de comprendre les façons dont les personnes font l'expérience des réalisations d'une politique publique et des types différents de résultats (intentionnels et non intentionnels) qu'elle est susceptible de produire à court, moyen et long terme, dans différents contextes macro, méso et micro.
- de développer et de pré-tester d'autres instruments de collecte de données qualitatives et quantitatives tels que des entretiens, des dispositifs expérimentaux ou des questionnaires.

III. Deux exemples d'utilisation de cette méthode : création d'indicateurs et évaluation de la mise en œuvre d'un programme de développement de l'enfant

Les *focus groups* devraient être utilisés dans l'évaluation des politiques publiques en fonction du type de preuves empiriques que l'on souhaite générer. Combinés à d'autres méthodes de production et d'analyse des données, ils sont souvent utilisés pour développer des indicateurs (par exemple, les taux de participation, l'incidence) qui peuvent aider à répondre aux questions évaluatives en caractérisant les réalisations et les résultats d'une politique d'une manière spécifique et mesurable. L'implication des bénéficiaires et d'autres parties prenantes dans le développement d'indicateurs peut rendre la politique pertinente pour eux et renforcer l'adhésion aux résultats de l'évaluation. EVALSED (Commission européenne, 2008), une ressource fournissant des conseils pour l'évaluation des politiques de développement socio-économique dans l'Union européenne, fournit un exemple d'utilisation de *focus groups* avec les bénéficiaires de la politique (par exemple, les responsables d'entreprises de la région) pour développer des indicateurs d'évaluation dans une politique de développement économique à Benton Harbour (Michigan, États-Unis).

Les évaluations formatives, qui visent à développer des politiques en examinant leur mise en œuvre, utilisent souvent des *focus groups*. L'évaluation formative du projet de développement de la petite enfance (DPE) de l'UNICEF dans le cadre du programme intégré de santé et de développement de la mère et de l'enfant (2017-2020) en Chine (Zhou Hong et al. 2022), a utilisé une méthode mixte fondée sur la théorie et axée sur l'utilisation, qui comprenait des *focus groups*. Les résultats de l'évaluation ont fourni des données probantes pour plaider en faveur de l'extension du modèle de développement de la petite enfance au niveau national et ont permis de concevoir le programme de développement de la petite enfance 2021-2025 de la Commission nationale de la santé et de

l'UNICEF. Les *focus groups* avec les jeunes parents et avec les soignant·e·s ont identifié les besoins en matière de compétences de *care* et ce résultat a été un élément moteur pour recommander la mise à l'échelle du DPE. La stigmatisation liée aux visites à domicile a également été pointée par certains *focus groups* et une attention supplémentaire à la protection de la vie privée a été recommandée dans le cadre de la mise à l'échelle. Les *focus groups* avec les administratrices et administrateurs ont renforcé les recommandations visant à augmenter le financement pour la mise en œuvre de trois types de services, à garantir la fréquence des services et à accroître leur couverture.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Étant donné qu'il existe de multiples approches d'évaluation qui diffèrent grandement dans leurs prémisses philosophiques et méthodologiques, il n'existe pas d'ensemble unique d'indicateurs de qualité pour la conduite des *focus groups* en évaluation. En effet, chacune de ces approches repose sur des hypothèses diverses et contradictoires et ce qui importe en termes de « qualité » varie en fonction de ces hypothèses.

De même, la recherche qualitative n'est pas une approche uniforme, elle englobe des traditions différentes basées sur des paradigmes différents, avec des hypothèses philosophiques diverses, qu'un cadre de qualité unique ne pourrait pas résumer. Le domaine des « critères de qualité des données qualitatives » est controversé, avec diverses positions et de nombreuses suggestions de classification disponibles, qui vont d'un rejet total de la notion de critères, à la promotion de critères similaires pour la recherche quantitative et qualitative.

Par conséquent, bien qu'il existe de nombreux critères de qualité sur le moment d'utiliser, la manière de concevoir, de recruter, de conduire et d'analyser les *focus groups*, il n'existe pas de normes convenues pour juger de la qualité des évaluations fondées sur des recherches qualitatives. Les *focus groups* se présentent sous de nombreux formats, et c'est pourquoi les questions pratiques, de conception et de qualité peuvent revêtir des caractères plutôt contrastés. Par exemple, le choix du lieu est important pour les *focus groups* en présentiel (par opposition aux *focus groups* en réalité virtuelle), car le succès du recrutement peut dépendre de l'accessibilité du lieu et des aspects pratiques (frais de déplacement, rafraîchissements, enregistrement audio); les questions de durée et d'animation sont toujours importantes, mais elles seront portées à un autre niveau dans les formes de discussion médiatisées par ordinateur (synchrones ou asynchrones).

Spencer *et al.* (2003, 16) ont proposé un cadre général pour les indicateurs de qualité sur quatre méthodes qualitatives, y compris les *focus groups*. Ce cadre est basé sur quatre principes directeurs essentiels : 1) Contribuer à faire progresser des connaissances ou une compréhension plus larges; 2) Avoir une conception et une stratégie défendables qui visent à répondre aux questions d'évaluation données; 3) Être rigoureux par une collecte de données systématique et transparente, ainsi que par l'analyse et l'interprétation des données; 4) Être crédible en offrant des arguments justifiables, défendables et plausibles sur la signification des données générées.

Ryan *et al.* (2014) ont proposé que les équipes d'évaluation considèrent des questions fondamentales pour maximiser leur apprentissage lors de la conduite de *focus groups* et pour améliorer la crédibilité des preuves empiriques générées par les *focus groups*, telles que : « Les participant·e·s au *focus group* ont-ils établi un terrain d'entente dans la conversation ou ont-ils et elles principalement agi en tant qu'individus? »; « Quelle était la dynamique de pouvoir entre l'équipe de modération et les participant·e·s, à la fois en tant que groupe et en tant qu'individus? »; « Quelles étaient les relations entre les participant·e·s – collectives ou dominantes? ».

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Les points forts de l'utilisation des *focus groups* sont les suivants :

- Au sein d'un groupe, les personnes peuvent s'appuyer sur les réponses des autres et les contester, et trouver des idées auxquelles elles n'auraient peut-être pas pensé seules. Ce riche mélange de perspectives et de désaccords peut éclairer les chercheuses et chercheurs sur les complexités des politiques, ce qui n'est souvent pas possible avec des méthodes moins dynamiques.
- Le format flexible est propice à l'exploration des résultats inattendus et des différences contextuelles.
- Les *focus groups* sont souvent recommandés comme une méthode permettant de gagner du temps et d'être rentable, mais les preuves de ces affirmations ne sont pas claires.

Les *focus groups* présentent les limites suivantes :

- Il est préférable de les utiliser dans le cadre d'une combinaison de méthodes plutôt que de façon isolée.
- Ils ne conviennent pas à la discussion de sujets trop sensibles et/ou controversés, car les gens sont moins susceptibles de s'ouvrir à ce sujet dans un groupe et peuvent avoir des craintes quant à la confidentialité et à l'anonymat.
- Souvent, ils ne permettent pas d'obtenir un niveau élevé de nuances ou de détails.

- Les *focus groups* peuvent être un défi pour les personnes ayant des difficultés de communication. Des stratégies d'inclusion pour toutes les capacités sont nécessaires, comme le choix de lieux et de salles accessibles, la conduite de discussions en ligne, des *focus groups* de plus petite taille, etc.
- Des *focus groups* attentifs aux différences culturelles doivent tenir compte non seulement de la langue et des identités culturelles, mais dans certaines cultures, il est préférable de les remplacer par d'autres méthodes décolonisées de conversation de groupe, comme les cercles de partage basés sur des récits à structure ouverte.
- Les *focus groups* ont une composition et une dynamique imprévisibles. Certains groupes de parties prenantes sont notamment difficiles à recruter pour les discussions de groupe.
- Les personnes qui sont à l'aise pour parler devant un groupe ont plus de chances d'être recrutées.
- Les discussions peuvent être détournées et/ou dominées par des individus ou des leaders d'opinion qui se font entendre. Les différences de statut entre les chercheurs et chercheuses et les participant·e·s, ou entre les participant·e·s, influencent les discussions.

Quelques références bibliographiques pour aller plus loin

Cohen-Miller, Anna. et Durrani, Naureen. et Kataeva, Zumrad. et Makhmetova, Zhadyra. 2022. « Conducting Focus Groups in Multicultural Educational Contexts: Lessons Learned and

- Methodological Insights. » *International Journal of Qualitative Methods*, 21: 1-10. <https://op.europa.eu/en/publication-detail/-/publication/752dd092-3ea5-429c-96dc-4507d0cda886>.
- European Commission. 2008. « EVALSED: The Resource for the Evaluation of Socio-Economic Development. » <https://op.europa.eu/en/publication-detail/-/publication/752dd092-3ea5-429c-96dc-4507d0cda886>.
- Krueger, Richard. et Casey, Mary Anne. 2000. *Focus Group: A Practical Guide for Applied Research*. Thousand Oaks, California: Sage.
- Manzano, Ana. 2022. Conducting focus groups in realist evaluation. *Evaluation*, 28(4): 406-425.
- Ryan, Katherine. et Gandha, Tyszha. et Culbertson, Michael. et Carlson, Crystal. 2014. « Focus Group Evidence: Implications for Design and Analysis. » *American Journal of Evaluation* 35 (3): 328-45.
- Spencer, Liz. et Ritchie, Jane. et Lewis, Jane. et Dillon, Lucy. 2003. « Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence. » London. <https://www.cebma.org/wp-content/uploads/Spencer-Quality-in-qualitative-evaluation.pdf>.
- Zhou, Hong. et Yang, Li. et Wang, Yan. et Liu, Shuang. et He, Hong. 2022. « Formative Evaluation of the National Health Commission-UNICEF Early Childhood Development Project of the Integrated Maternal and Child Health and Development Programme (2017-2020). » <https://www.unicef.org/evaluation/reports#/detail/13964/formative-evaluation-of-the-national-health-commission-unicef-early-childhood-development-project-of-the-integrated-maternal-and-child-health-and-development-programme-2017-2020>.

12. Entretiens de groupe

CHARLOTTE HALPERN

Résumé

L'entretien de groupe¹ est une méthode qualitative par laquelle un entretien semi-directif est mené avec plusieurs personnes en même temps. Cette méthode a pour résultat de créer artificiellement un ensemble d'interactions sociales entre un nombre sélectionné de participant·e·s, par exemple différent·e·s acteur·rice·s de la politique. Elle est utile à diverses formes d'évaluation des politiques publiques, notamment l'évaluation *ex ante*, *ex post* et de processus.

Mots-clés : Méthodes qualitatives, entretien, entretien de groupe, élites, pluralité, constructivisme, interprétativisme

1. Mon intérêt pour les entretiens de groupe en tant que méthode de recherche en politiques publiques découle de l'expérience accumulée dans le cadre d'un travail sur les politiques de transition vers la mobilité durable dans les villes européennes. Entre 2015-2022, j'ai organisé une vingtaine d'entretiens de groupe dans 14 villes et à Bruxelles avec une variété d'acteur·rice·s. Je suis particulièrement reconnaissante aux partenaires du projet H2020 CREATE (subvention n°636573), en particulier Peter Jones, Charles Buckingham et Lucia Cristea, pour avoir soutenu l'idée de tester cette méthode pour analyser le rôle des politiques publiques visant à limiter l'usage de la voiture; aux partenaires du projet H2020 MORE (subvention n°769276) qui a fourni l'opportunité de renforcer la méthodologie à partir de la suggestion de Jenny McArthur de lier les entretiens de groupe à un exercice de cartographie des parties prenantes et de les intégrer dans l'analyse textuelle du corpus de données collectées, et enfin, les partenaires du projet H2020 CIVITAS SUMP PLUS (subvention n°814881) au cours duquel j'ai expérimenté des entretiens de groupe hybrides et à distance en raison de la pandémie COVID-19.

I. En quoi consiste cette méthode?

Les entretiens menés avec plusieurs personnes en même temps font référence à un ensemble diversifié de méthodes qualitatives bien connues dans la recherche en sciences sociales. Leur utilisation spécifique dépend du rôle et de la fonction qu'ils occupent dans une stratégie de recherche (Knott et al., 2022), ainsi que de leurs propriétés (Duchesne et Haegel, 2008). Parmi eux, les entretiens de groupe présentent un intérêt particulier pour la recherche en évaluation des politiques publiques. Ils ne doivent pas être confondus avec d'autres techniques telles que les discussions de groupe, les focus groups et les pré-tests, principalement parce qu'ils ne nécessitent pas d'être centrés sur une expérience commune, ni que les participant·e·s partagent des statuts professionnels et sociaux homogènes (Marier et al., 2020). En créant artificiellement un ensemble d'interactions sociales entre un nombre sélectionné de participant·e·s, ils diffèrent des méthodes ethnographiques, notamment des observations.

Les entretiens de groupe sont une technique utile pour lancer une discussion de groupe informelle avec un petit groupe d'acteur·rice·s bien informé·e·s et de haut niveau — parfois appelés « élites » (Glas, 2021) — dont la contribution est jugée pertinente pour la compréhension de la question étudiée, par exemple l'évaluation d'un programme de politique publique. La valeur ajoutée des entretiens de groupe ne réside pas dans le temps gagné en interrogeant plusieurs personnes en même temps — cette vision est erronée car les entretiens de groupe nécessitent un travail de préparation et de traitement des données tout aussi sinon plus important que les entretiens individuels (voir chapitre séparé sur les entretiens semi-directifs) — mais dans l'opportunité de générer artificiellement des interactions sociales entre une diversité d'acteur·rice·s. Il permet d'identifier et de donner du sens à une pluralité de perspectives, d'intérêts et de valeurs, ainsi que de mettre en lumière d'éventuelles contradictions et ambiguïtés. En plus de la relation entre le répondant ou la répondante et l'interview·eur·euse, l'évolution des

relations entre les membres du groupe peut être un stimulant pour l'élaboration et l'expression. Ainsi, les entretiens de groupe permettent de tirer parti de la dynamique de groupe pour produire des données nouvelles et supplémentaires (Frey, Fontana, 1991, 183).

Cette technique peut jouer un rôle décisif dans un design de recherche qualitative, et ce de différentes manières. Tout d'abord, lorsqu'ils sont introduits dans une perspective exploratoire au tout début de la recherche, les entretiens de groupe sont particulièrement utiles dans le cas d'un sujet peu étudié, pour lequel les sources sont rares et insuffisamment diversifiées. Deuxièmement, en s'appuyant sur un « effet de groupe », les interactions peuvent favoriser l'émergence de perspectives différenciées sur un sujet donné, qui n'auraient pas pu être saisies par le biais d'observations ou dans le cadre d'entretiens individuels. En tant que tels, les entretiens de groupe permettent de générer artificiellement un ensemble d'interactions sociales pour exprimer des points de vue partagés ou des désaccords sur un sujet donné (Morgan, 1997), tout en laissant la possibilité de réaliser d'autres entretiens individuels avec un nombre réduit de participant·e·s. Troisièmement, pour les praticien·ne·s situé·e·s au sommet de leur structure organisationnelle, participer à une discussion de groupe constitue un facteur décisif pour trouver le temps de l'entretien (Glas, 2021). Quatrièmement, au tout début de la recherche, ils peuvent être utilisés pour examiner la robustesse de l'ensemble des hypothèses résultant de l'analyse documentaire et pour les affiner en conséquence. Ainsi, les entretiens de groupe sont également pertinents dans le cadre d'une recherche comparative, le même guide d'entretien pouvant être appliqué à l'ensemble des cas étudiés pour fournir un premier aperçu comparatif général et générer quelques hypothèses spécifiques à chaque cas.

Cette technique repose sur la sélection de 8 à 12 participant·e·s, visant à réunir un ensemble d'acteur·rice·s de premier plan, représentant une diversité de points de vue sur l'objet de l'étude en raison de leurs antécédents, rôles et fonctions respectifs dans leurs propres organisations. Le degré de diversité peut varier en fonction du contexte

politique et de la question de recherche. Ainsi, la priorité peut être donnée au moment de la sélection, à une variété de formations et d'antécédents professionnels pour garantir des échanges interdisciplinaires, à une variété de rôles et de fonctions² pour tenir compte des points de vue, de contextes et de priorités les plus divers possibles, ou encore, pour refléter un large éventail d'organisations dans un contexte de politique publique donné. Dans une recherche qui couvre une période longue de 40 à 50 ans, la sélection des participant·e·s peut viser la diversité générationnelle.

En fonction de la question de recherche évaluative, de la disponibilité des sources et de la langue de collecte de ces données, ce corpus de données qualitatives peut être codé pour faire l'objet d'une analyse à l'aide d'un logiciel d'analyse qualitative tel que InVivo (voir également Knott et al., 2022). Il peut ainsi faire l'objet d'une analyse de texte ou de discours, mais aussi pour produire une cartographie d'acteurs ou une chronologie des politiques publiques, contribuant ainsi à la constitution d'un socle solide en préalable à un questionnement évaluatif plus ciblé.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Les apports des entretiens de groupe à la recherche qualitative ont déjà été abordés. Pour les recherches sur l'évaluation des politiques publiques, cette technique offre la possibilité de réexaminer les frontières de problèmes politiques bien connus ainsi que les relations causales les mieux établies (Zittoun et al., 2021). Dans une perspective constructiviste et interprétative, cette méthode adopte un point de vue critique sur les prémisses rationalistes de l'analyse des politiques publiques et met en

2. Élu·e, technicien·ne, fonctionnaire, militant·e d'une ONG, chef·fe d'entreprise, etc.

évidence les contraintes pesant sur les activités d'évaluation (Wollman, 2006). Elle reconnaît que les objectifs des politiques publiques (en tant que conséquences souhaitées) sont souvent vagues, ambigus, potentiellement contradictoires ou mutuellement exclusifs. Ces mêmes objectifs sont par ailleurs compris de différentes manières par les principaux acteurs de la fabrique des politiques publiques, sans parler des publics cibles et des bénéficiaires, et bien qu'elles ne soient pas nécessairement exactes, ces diverses compréhensions des problèmes publics et des solutions de politiques publiques sont réinjectées dans le processus politique, influençant son pilotage et ses développements futurs. Cela soulève d'importants problèmes de causalité, surtout pour les politiques publiques caractérisées par un degré élevé de complexité et d'incertitude, et en période de crise (Voss et Kemp, 2006). Sur la base de ces observations, les entretiens de groupe cherchent ainsi à générer artificiellement un ensemble d'interactions sociales permettant d'examiner de manière critique les relations causales entre, d'une part, les changements attendus ou observés, et d'autre part, un programme ou une mesure politique donnée.

Dans cette optique, les entretiens de groupe sont utiles pour une variété de questions évaluatives, telles que l'évaluation *ex ante*, *ex post* et de processus, que ce soit en combinaison avec d'autres méthodes d'évaluation ou de manière autonome. En ce qui concerne l'évaluation *ex ante*, elle peut être utilisée comme une occasion d'examiner les relations causales (plus ou moins) explicites entre les objectifs déclarés, la sélection des moyens, ainsi que les résultats attendus (voir chapitre séparé sur l'évaluation basée sur la théorie). En outre, elle permet de comprendre comment différentes options de politiques alternatives sont débattues, quelles visions du monde et quels arguments sont utilisés, et quelles stratégies d'atténuation des risques sont développées pour surmonter les résistances attendues. Cela peut, à son tour, alimenter la prise de décision et mettre en lumière les contradictions et ambiguïtés existantes. Les entretiens de groupe se sont également avérés particulièrement utiles pour alimenter l'évaluation des processus, que ce soit en mode

accompagnement (en parallèle) ou en mode intervention. Dans ce cas, sa fonction est d'identifier les facteurs intermédiaires intervenant pendant la mise en œuvre. Enfin, dans le cas des évaluations *ex post*, qu'elles soient axées sur les méthodes ou les résultats, les entretiens de groupe apportent un éclairage complémentaire aux questions évaluatives ciblées, contribuant souvent à faire sens d'une éventuelle déconnexion entre les objectifs annoncés des politiques publiques et leurs effets non prévus, à discuter de l'utilisation d'un ensemble d'indicateurs et à susciter un débat sur les futurs programmes de politiques publiques. Ceci, à son tour, peut contribuer à identifier d'éventuels processus d'apprentissage, en tant qu'objet de recherche ou d'intervention.

III. Un exemple d'utilisation de cette méthode

Les entretiens de groupe ont été utilisés dans divers contextes de recherche en politiques publiques, notamment pour des questions d'évaluation. Dans le contexte de la crise climatique, elle ouvre de nouvelles voies pour l'évaluation des politiques de transition et d'adaptation. En tant que politiques visant à atteindre des objectifs à long terme, les politiques de transition et d'adaptation font référence au passage du possible au souhaitable, et les progrès sont évalués par rapport à des futurs politiques qui ne sont pas sans équivoque. Contrairement aux problèmes techniquement clairs, les problèmes des politiques de transition ne reposent pas sur une définition ou une solution claire, ils sont caractérisés par des relations causales incertaines et ils rassemblent un large éventail de parties prenantes ayant des valeurs ou des intérêts contradictoires, ce qui explique les désaccords constants sur les moyens de résoudre les problèmes (Van der Steen et al., 2016). Cela favorise la nécessité de s'appuyer sur des modèles de recherche évaluative dans lesquels les degrés de divergence des valeurs sont délibérément examinés et débattus (Delahais et al., 2020).

En se concentrant sur les transitions de mobilité durable, Hickman et Banister (2014) ont examiné dans quelle mesure le futur constitue un défi pour les décideur·euse·s politiques, ainsi que les lacunes des méthodes dominantes telles qu'identifiées dans la littérature, comme la prévision et la modélisation en particulier, ou les approches classiques utilisées dans l'analyse de scénarios. S'appuyant sur les résultats du projet Urban Buzz³, ils expliquent comment une approche rétrospective de la planification des transports à Londres a été mise en place dans le but explicite d'évaluer l'efficacité carbone de la stratégie existante et de contribuer au développement d'une nouvelle stratégie visant à réduire de 60% les émissions carbone dans le secteur des transports d'ici 2025 et 2050. La recherche s'est appuyée sur une combinaison de méthodes, notamment des entretiens de groupe, qui ont pris la forme d'ateliers avec des décideur·euse·s et, alternativement, avec une diversité d'acteur·rice·s, et ce, à chaque étape du processus. Le design de recherche visait explicitement à ramener le rôle des valeurs dans le cadre analytique, à évaluer la diversité des représentations sur les futurs de la transition, la hiérarchie des valeurs associées aux processus de transition, l'éventail des stratégies de mise en œuvre et d'interroger chacun de ces choix. En contribuant au développement d'un scénario rétrospectif étroitement articulé avec une stratégie de mise en œuvre, le projet a confirmé la pertinence de l'examen des valeurs des acteurs en présence pour aborder les visions du futur dans le secteur des transport et, ainsi, contribuer à l'émergence d'une approche distincte de la mobilité à Londres.

3. Voir le projet VIBAT London (Looking Over the Horizon: Transport and Global Warming - Visioning and Backcasting for Transport in London) sur le site Web du projet : https://www.ucl.ac.uk/urbanbuzz/projects_28.php (dernière consultation le 8 novembre 2022).

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Il est erroné de croire que l'*interview* simultanée d'une diversité d'acteur·rice·s permettrait de gagner du temps. Cette technique nécessite un important travail préparatoire et d'analyse des données (Duchesne et Hagel, 2008). De nature exploratoire, les entretiens de groupe sont, en effet, fondés sur une recherche préliminaire approfondie, telle qu'une revue de la littérature, une évaluation de la disponibilité des données – littérature grise, rapports publics, coupures de presse, manifestes de partis politiques, etc. – et une cartographie des principaux·ales acteur·rice·s. Ce travail préalable alimente la production d'un guide d'entretien, qui contribue à structurer la discussion tout en préservant la nature exploratoire de la démarche de recherche. Il peut inclure un petit nombre de questions ciblées pour guider la discussion. En complément, des discussions en petits groupes peuvent être encouragées dans le cadre de séquences dédiées, pour encourager des échanges approfondis sur un sujet donné ou générer, de manière collective, une chronologie précise et partagée de l'évolution d'une politique publique dans la durée.

L'organisateur de l'entretien de groupe doit également être conscient des difficultés consistant à réunir un groupe aussi diversifié d'acteur·rice·s, surtout si le sujet fait l'objet de controverses. Tout en cherchant à favoriser une discussion informelle et animée, les entretiens de groupe doivent se dérouler dans un cadre préétabli. Par ailleurs, les participant·e·s peuvent être réticent·e·s à participer à un entretien de groupe, craignant qu'il ne débouche sur des échanges superficiels. Il est donc essentiel de le présenter clairement comme une méthode de recherche et de fournir une structure (légère) pour éviter les discussions trop générales et futiles. L'entretien ne doit pas durer plus de 3-4 heures, mais prévoir une pause peut contribuer à structurer la discussion collective tout en offrant la possibilité d'échanges plus informels. Il est d'ailleurs fréquent que les participant·e·s s'en fassent l'écho lors de la

reprise de l'entretien. Pour éviter de mettre les participant·e·s dans une position difficile, celles et ceux-ci doivent être informé·e·s à l'avance des principales caractéristiques de l'entretien et de la liste des participant·e·s. Ils et elles doivent aussi donner leur consentement éclairé. Les décisions relatives à l'anonymat ou à la confidentialité, au stockage et à la diffusion des données, doivent être abordées lors de la demande du consentement éclairé des participants, qu'il soit écrit ou oral. En fonction de l'approche choisie pour l'analyse des données, les entretiens de groupe peuvent faire l'objet d'un enregistrement audio et des notes détaillées peuvent être prises pendant la discussion pour les besoins de l'équipe de recherche. Aucun public extérieur aux organisateur·rice·s des entretiens et aux participant·e·s ne doit être admis.

Les entretiens de groupe nécessitent donc un important travail préparatoire pour décider de la sélection des participant·e·s, du guide d'entretien et de l'utilité de prévoir des discussions en petits groupes pour approfondir une question spécifique.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Pour conclure, les entretiens de groupe présentent plusieurs avantages pour la recherche et la pratique de l'évaluation des politiques publiques. Lorsqu'ils sont utilisés dans une perspective exploratoire, au tout début de la recherche, ils permettent d'examiner la robustesse des hypothèses résultant de l'analyse documentaire, de fournir un premier aperçu comparatif général et de générer des hypothèses spécifiques au contexte. En générant artificiellement un ensemble d'interactions sociales ou « effet de groupe », ils donnent l'occasion aux participant·e·s d'exprimer des points de vue partagés ou des désaccords sur un sujet donné. En tant

que tel, il s'agit d'une technique puissante de collecte de données, qui permet de jeter un regard neuf sur un sujet donné, peu accessible par le biais d'observations ou dans le cadre d'entretiens individuels.

En générant artificiellement un ensemble d'interactions, « l'effet de groupe » produit un ensemble de données très original, composé de nouvelles informations et éléments de preuve. En partageant leurs points de vue et leurs désaccords potentiels sur un enjeu, un programme ou une solution de politique publique, sa chronologie, ses relations causales et ses effets sont remis en discussion, contribuant ainsi à ouvrir de nouvelles perspectives pour la recherche évaluative ou à informer la fabrique des politiques publiques. En outre, les entretiens de groupe permettent de générer un ensemble solide d'hypothèses générales et au cas par cas, de remettre en question la pertinence des facteurs, internes et externes, de changement, d'identifier les effets d'une mesure politique donnée tout en tenant compte de considérations politiques plus larges (et en remettant en question ses effets (involontaires)).

À l'inverse, cette technique est mal adaptée à un questionnement évaluatif ciblé. D'autres méthodes qualitatives, telles que les *focus groups*, seraient mieux adaptées, principalement parce que les entretiens de groupe n'exigent pas que les participant-e-s partagent une expérience commune, des statuts professionnels et sociaux homogènes. De plus, les entretiens de groupe cherchent à créer artificiellement un ensemble d'interactions sociales entre un nombre sélectionné de participant-e-s encouragé-e-s à exprimer leurs désaccords sur un sujet donné, qu'il s'agisse du diagnostic du problème, de la hiérarchie des valeurs pour choisir une ligne de conduite ou de ses effets. En tant que telles, cette technique diffèrent également des méthodes ethnographiques, notamment les observations, mais aussi des entretiens individuels.

Quelques références bibliographiques pour aller plus loin

- Delahais, Thomas. et Sage, Kate. et Honoré, Vincent. 2020. Evaluators in Transition. *Zeitschrift für Evaluation (ZfEv)*. 2: 239-260.
- Duchesne, Sophie. et Haegel, Florence. 2008. *L'enquête et ses méthodes: l'entretien collectif*. Paris: Armand Colin.
- Frey, James H.. et Fontana, Andrea. 1991. The group interview in social science research. *Social Science Journal*. 28(2): 175-187.
- Glas, Aarie. 2021. Positionality, power and positions of power: reflexivity in elite interviewing. *PS, Political science & politics*, 54(3): 438-442.
- Hickman, Robin. et Banister, David. 2014. *Transport, climate change and the city*. London: Routledge.
- Knoll, Eleanor. et Hamid Rao, Aliya. et Summers, Kate. et Teegeer, Chana. 2022. Interviews in the social sciences. *Nat Rev Methods Primers*. 2(73). <https://doi.org/10.1038/s43586-022-00150-6>
- Marier, Patrick. et Dickson, Daniel. et Dubé, Anne-Sophie. 2020. Using focus groups in comparative policy analysis. In Peters, B. Guy, et Fontaine, Guillaume. Eds. *Handbook of Research Methods and Applications in Comparative Policy Analysis*. Cheltenham: Edward Elgar Publishing: 297-310.
- Morgan, David L. 1997. *Focus Groups as Qualitative Research*, London: Sage.
- Steen, Martijn van der. et Chin-A-Fat, Nancy. et Vink, Martinus. et Twist, Mark van. 2016. Puzzling, powering and perpetuating: Long-term decision-making by the Dutch Delta Committee. *Futures*. 76: 7-17.

Voß, Jan-Peter. et Kemp, René. 2006. Sustainability and reflexive governance: introduction. In Voß, Jan-Peter. et Bauknecht, Dieter. et Kemp, René. Eds. *Reflexive Governance for Sustainable Development*. Cheltenham: Edward Elgar.

Wollmann, Hellmut. 2006. Evaluation and evaluation research. In Fischer, Frank, Miller, Gerald J., Sidney Mara S., *Handbook of Public policy analysis*. London: Routledge.

13. Les études de cas

VALÉRY RIDDE, ABDOURAHMANE COULIBALY, LARA GAUTIER

Résumé

Les études de cas consistent à analyser de façon approfondie un ou plusieurs cas, à partir d'une diversité de méthodes et au regard d'approches théoriques. Le choix des cas (unique ou multiples) étudiés est crucial. Les études de cas sont particulièrement adaptées pour étudier l'émergence et les processus en jeu dans la mise en œuvre des politiques et pour participer aux évaluations basées sur la théorie.

Mots-clés : Méthodes qualitatives, méthodes quantitatives, méthodes mixtes, étude de cas, approches théoriques, cas unique/multiples, triangulation empirique, généralisation analytique

I. En quoi consiste cette méthode?

Aussi employée en anthropologie, la démarche d'étude de cas a été utilisée depuis longtemps en évaluation, où elle n'est pas considérée comme une méthode mais comme une stratégie de recherche (Yin 2018). En étudiant une politique dans son contexte et au moyen de multiples sources de données, l'étude de cas (unique ou multiples) cherche à répondre à des questions du type « comment » et « pourquoi » à partir d'une démarche systémique et avec les soutiens d'approches théoriques. La réalisation d'une étude de cas pour l'évaluation d'une politique publique suit un processus habituel en évaluation : planification, rédaction du protocole, préparation du terrain, collecte et analyse des données, partage des résultats et éventuelles recommandations pour

l'amélioration de la politique (Gagnon 2012). Comme pour toutes les évaluations, le choix des méthodes doit suivre les objectifs et la question d'évaluation, et non pas l'inverse. Une étude de cas pourra ainsi mobiliser des méthodes qualitatives, quantitatives et différents devis (*design*) de méthodes mixtes.

La stratégie des études de cas est donc appropriée lorsqu'il s'agit d'organiser une évaluation de l'émergence, des processus, de la pertinence ou de l'adaptation des politiques. Elle est souvent mobilisée lorsque les équipes évaluatives disposent de peu ou d'aucun contrôle sur les événements et le contexte qui influence les actions de la politique. Cela est bien souvent le cas en dehors de toutes situations expérimentales, rares dans le domaine des politiques publiques. Elle est donc surtout préconisée pour comprendre un phénomène contemporain, souvent complexe, et organisé dans un contexte réel.

Le recours à la démarche des études de cas peut servir à expliquer une politique publique, la décrire en profondeur ou encore illustrer une situation spécifique, qui parfois peut être originale et éclairante pour les prises de décision. L'intérêt des études de cas est de pouvoir s'adapter à des situations différentes pour lesquelles il existe de multiples variables d'intérêts autour d'une politique. Il s'agit aussi de pouvoir utiliser de multiples sources de données, quantitatives ou qualitatives, qui permettent d'assurer une triangulation empirique. La stratégie des études de cas permet de tenir compte de propositions théoriques et de l'état des connaissances scientifiques pour orienter la collecte et l'analyse des données. Elle s'inscrit parfaitement, sans s'y limiter, dans les démarches d'évaluation basées sur la théorie.

Il existe une myriade de propositions pour qualifier les types d'études de cas possibles. D'abord, il est possible d'utiliser des études de cas simples/ uniques (concernant une politique) ou des études de cas multiples (plusieurs politiques dans un même contexte organisationnel ou une même politique dans des contextes différents). Ensuite, ces cas peuvent être étudiés de manière holistique (la politique dans sa globalité) ou selon

différents niveaux d'analyses (les dimensions de la politique que la théorie d'intervention aura précisées ou les contextes régionaux particuliers). Le choix des cas étudiés doit être heuristique (pouvoir apprendre de l'étude) et stratégique (pouvoir disposer de données selon le budget disponible, répondre à des questions utiles). L'un des critères essentiels du choix des cas est de pouvoir disposer d'informations suffisamment pertinentes pour comprendre la politique en profondeur et dans sa complexité. L'échantillonnage des cas doit donc être explicite, rigoureux et transparent. Le choix des études de cas peut ainsi permettre de disposer de cas critiques, uniques, typiques, révélateurs, instrumentaux, etc. Cette sélection peut aussi être réalisée en collaboration entre les équipes de recherche et celles qui pilotent la politique afin de s'assurer de la pertinence des choix mais aussi de leur faisabilité. La sélection peut aussi reposer sur des analyses quantitatives préalables pour disposer de la situation de départ des cas et, par exemple, choisir des cas très contrastés ou très similaires dans leur performance à l'égard de la politique analysée.

Parfois, il peut aussi s'avérer fécond de disposer d'une approche diachronique afin de produire des études de cas longitudinales. Ainsi, l'analyse dans le temps d'une politique permet de mettre au jour les influences de l'évolution du contexte ou des stratégies des personnes qui la mettent en œuvre, ou encore de celles en bénéficiant. Partir de cas dont les conditions initiales sont semblables pour ensuite étudier leur évolution est qualifié de « *racing cases* » par Eisenhardt (Gehman *et al.* 2018).

Lors de l'analyse des données, la démarche de l'étude de cas nécessite, en plus des analyses habituelles spécifiques aux méthodes (analyse de contenu, thématique, statistiques descriptives ou inférentielles, etc.) de mobiliser une logique de réplication. L'idée est de comparer, de manière systématique et rigoureuse, les données empiriques et la théorie, qu'elle soit la théorie d'intervention de la politique ou un cadre théorique ou conceptuel utilisé pour comprendre la politique. Ce processus est nommé

par Yin la *généralisation analytique*. Lorsque plusieurs cas soutiennent la même théorie, il est possible de suggérer la présence d'une logique de réplication (Yin 2010).

Les configurations peuvent être des outils heuristiques pour cette démarche d'analyse, qu'elles soient organisationnelles ou ancrées dans le réalisme critique (voir le chapitre séparé sur l'évaluation réaliste). En outre, retrouver des configurations, ou des situations semblables dans des contextes différents, renforce la capacité à généraliser les résultats des études de cas. Yin estime que la généralisation analytique nécessite la construction d'un argumentaire très solide et qui sera en mesure de résister aux défis des analyses logiques. Ainsi, il est indispensable de préciser cet argumentaire théorique dès le début de l'étude de cas, soit en mobilisant une théorie ou à partir de l'état des connaissances sans que cela soit parfaitement spécifique à la politique publique analysée. Au départ d'une étude de cas, il faut donc rester à un niveau conceptuel relativement élevé, à tout le moins supérieur à la politique étudiée. Ensuite, les résultats empiriques de l'étude de cas doivent montrer en quoi ils s'alignent (ou pas) avec l'argumentaire théorique de départ. Il faudra enfin discuter de la manière dont cette réflexion théorique, à partir de cette politique particulière, peut aussi s'appliquer à d'autres situations et d'autres politiques de l'étude de cas particulière. Le fait d'avoir, y compris au début de l'étude de cas, formulé également un contre-argument (des « *hypothèses rivales* ») et d'avoir tenté de disposer de données empiriques au cours de la collecte de données (qui les réfutent), renforce la validité de ce processus de généralisation analytique. Enfin, la puissance des études de cas multiples est que cette généralisation analytique est renforcée lorsque les résultats d'un cas sont similaires à ceux des autres cas.

Certaines équipes de recherche proposent même que les études de cas puissent conduire à la construction de théories (*theory-building case studies*), notamment lorsque l'on analyse des objets complexes comme les politiques publiques.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Avant de décider de s'engager dans la voie de l'étude de cas, il convient de se poser deux questions préliminaires qui détermineront le bien-fondé de l'approche :

1. Est-ce que le phénomène auquel je m'intéresse a besoin du (ou des) cas pour être compréhensible? (p.ex., *Theory-building case studies*)
2. Est-ce que le(s) cas représente(nt) une fenêtre empirique qui éclaire l'analyse du phénomène plus large?

Une fois que l'on a répondu positivement à l'une ou à l'autre, on peut définir les questions évaluatives:

- Dans quelles conditions de la vie réelle la politique publique X expérimentée sous forme de pilote dans le contexte A, peut-elle être mise à l'échelle dans les contextes B, C, et D?
- Comment la controverse au sujet de la politique publique Y apparue dans le contexte B a-t-elle émergé?
- Quels sont les facteurs de succès de la mise en œuvre de la politique publique X dans le contexte A?
- Comment les politiques publiques Y et Z ont-elles été mises en œuvre dans le contexte B ?
- Pourquoi la politique publique X dans les contextes A et B a-t-elle échoué, alors qu'elle a eu des effets positifs dans le contexte C?
- Pourquoi la politique publique X mise en œuvre dans le contexte A a échoué, alors que la politique publique Y mise en œuvre dans le même contexte A a réussi?

- Qu'est-ce qui, dans les caractéristiques de la politique publique Z mise en œuvre dans les contextes A, B, et C, permet d'informer la théorie μ ? (*Theory-building case studies*)

L'usage de l'étude de cas peut intervenir à n'importe quel moment du processus d'évaluation, *ex ante* (au moment de la conception de la politique), *in itinere* (pendant la mise en œuvre), ou *ex post* (par exemple, afin de mieux comprendre les résultats produits).

III. Un exemple d'utilisation de cette méthode au Burkina Faso

Des études de cas simples et multiples longitudinales ont été mobilisées pour étudier une politique publique de financement de la santé au Burkina Faso (Ridde 2021).

La Banque mondiale a incité le gouvernement à tester dans une douzaine de districts une modalité de financement des centres de santé supplémentaire au budget de l'État. Il s'agissait d'organiser un paiement à la performance où les centres de santé et les professionnels et professionnelles de santé recevaient des fonds supplémentaires en fonction de l'atteinte de résultats d'activités. Par exemple, pour chaque accouchement réalisé dans le centre avec un partographe, ils recevaient 3,2 euros à partager entre la structure et le personnel, selon des procédures et des indicateurs complexes. Des processus de vérification et de contrôle étaient organisés pour s'assurer de la fiabilité des demandes de paiements.

Pour étudier l'émergence de cette nouvelle politique, nous avons réalisé une étude de cas unique (centrée sur la politique) afin de mieux comprendre son origine, les idées véhiculées, les solutions proposées, les personnes l'ayant proposée, les enjeux de pouvoir, etc. Nous avons employé une analyse documentaire et 14 entretiens qualitatifs en

profondeur auprès des responsables politiques, des organismes de financements et des personnes expertes du sujet. Selon une démarche de généralisation analytique, nous avons comparé cette émergence pour comprendre si ce qui s'est déroulé au Burkina Faso se reproduisait aussi au Bénin.

Pour étudier la mise en œuvre de la politique au Burkina Faso, nous avons eu ensuite recours à des études de cas multiples longitudinales. Pour des raisons de temps et de budget, nous avons sélectionné trois districts représentant la diversité des situations de mise en œuvre de la politique. Puis, au sein de chacun de ces districts, nous avons sélectionné six cas parmi les centres de santé primaires (environ 30 par district) et un cas qui était l'hôpital de référence (un seul par district). Les six cas ont été sélectionnés en fonction des trois types de stratégies de financement que la politique souhaitait tester, donc deux cas par type. Nous avons décidé de retenir deux cas les plus contrastés possibles au sein de chacun des trois types : un centre de santé très performant et un autre pas du tout. La performance a été calculée à l'aide d'une méthode quantitative (série chronologique) sur la base d'indicateurs de fréquentation des centres de santé au cours des années précédant la politique. Cette analyse *étiquée (du point de vue externe)* a classé tous les centres de santé selon leur ordre de performance pour soutenir la sélection des cas. Cette dernière a aussi bénéficié de l'avis *émique (du point de vue interne)* des responsables locaux du système de santé afin de tenir compte de leur propre perception de la performance des centres, au-delà de l'approche quantitative qui ne donne qu'une vision partielle de la performance. Ainsi, pour chacun des sept cas sélectionnés par district ($7 \times 3 = 21$), nous avons employé de multiples sources de données pour comprendre les défis de la mise en œuvre de la politique : analyse de la documentation, entrevues qualitatives formelles (entre 114 et 215 par district) et informelles (entre 26 et 168 par district), observations de situations. Une grille de collecte de données a aussi permis de mesurer la fidélité de la mise en œuvre de la politique. Afin de mieux comprendre l'évolution de la mise en œuvre

de la politique, et notamment les adaptations au cours du temps, trois moments de collecte de données ont été réalisés sur une période de 24 mois, suivant ainsi l'approche des études de cas multiples longitudinales.

Enfin, ces études de cas ont aussi été fécondes pour étudier, avec une approche qualitative et une longue immersion sur le terrain, les conséquences non attendues (positives ou négatives) de cette politique. Si cette dimension de l'évaluation est encore trop peu appréhendée, sa réalisation au Burkina Faso a montré toute la pertinence de cette démarche (Turcotte-Tremblay *et al.* 2017). Se limiter aux effets attendus, qu'implique souvent une focalisation extrême sur la seule théorie d'intervention développée par les équipes qui définissent la politique, réduit la portée heuristique de l'évaluation. Si les succès sont essentiels, les défis peuvent aussi être indispensables pour améliorer les politiques publiques à l'aide des études de cas.

Pour toutes ces démarches, l'analyse s'est réalisée de manière hybride, soit déductive (au regard de la théorie d'intervention ou d'un cadre conceptuel) et inductive (données empiriques originales). La comparaison entre les cas, entre les districts et entre les pays a permis de monter en abstraction dans une démarche de généralisation analytique.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Juger de la qualité d'une démarche complexe comme celle des études de cas nécessite une vision globale, en dépassant les réflexions spécifiques, mais essentielles, aux méthodes habituelles (quantitative et qualitative). Pour cela, Yin (2018) propose d'étudier la qualité des études de cas au regard des quatre dimensions :

- Validité de construit (étudier la politique attendue et pas autre chose) : utiliser de multiples sources de données probantes, décrire et établir une chaîne causale, impliquer les parties prenantes dans la validation du protocole et des rapports;
- Validité interne (confiance dans les résultats) : comparer les données empiriques entre elles et avec la théorie, construire des logiques d'explications, tenir compte des hypothèses rivales et différentes, recourir aux cadres logiques/théories de l'intervention;
- Validité externe (capacité de généraliser les résultats) : utiliser des théories, utiliser la logique de reproduction analytique;
- Fiabilité (pour la même étude de cas, les mêmes conclusions) : recourir à un protocole d'étude de la politique, développer une base de données des cas.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Le principal atout de l'étude de cas est sa capacité à « *intégrer les caractéristiques uniques de chaque cas et d'examiner des phénomènes complexes dans leur contexte* », c'est-à-dire, en condition de vie réelle (Stiles 2013, 30).

La stratégie de l'étude de cas, du fait de l'abondance et la variété des corpus de données mobilisés, et des méthodes de recherche employées (qualitatives, quantitatives ou mixtes), permet, le plus souvent, une riche description de la ou des politique(s) publique(s) évaluée(s) et des contextes de mise en œuvre. C'est notamment le cas des études de cas uniques, qui permettent d'aller en profondeur dans l'analyse. En ce qui concerne les études de cas multiples, l'avantage principal est que cela permet d'accroître les variations potentielles, ce qui augmente la

robustesse de l'explication. Le revers de la médaille est que ces stratégies requièrent de dédier un temps important. Ainsi, l'ampleur du travail peut s'avérer problématique, surtout si les délais fixés par les commanditaires sont courts. De plus, s'il y a plusieurs questions évaluatives, ou une question qui invite à relier les enjeux de mise en œuvre aux effets, il faudra alors, sans doute, penser à associer l'étude de cas (qui pourra se concentrer sur l'analyse de processus, par exemple) à une autre stratégie de recherche complémentaire, tel que les démarches quasi-expérimentales (Yin et Ridde 2012). Enfin, plusieurs biais peuvent apparaître; le choix orienté du ou des cas, la faible puissance statistique lorsque l'on réalise des analyses quantitatives. Ces biais sont susceptibles d'éroder la comparabilité inter-cas ou entre les contextes. La riche justification du choix des cas (les politiques publiques) (Stake 1995) et la description du ou des contextes, ainsi que le processus de généralisation analytique, décrit précédemment, permettent de diminuer l'impact de ces biais.

En ce qui concerne les *theory-building case studies*, on repère à la fois des avantages et des inconvénients de l'étude de cas (Stiles 2013). La stratégie de l'étude de cas consiste ici à comparer différents énoncés provenant de la théorie, avec une seule ou plusieurs observations. Ceci peut se réaliser à travers la description du ou des quelques cas en des termes théoriques. Ainsi, si chaque détail ne peut être observé qu'une seule fois, ils peuvent s'avérer très nombreux et donc utiles à la construction de la théorie. Toutefois, les mêmes biais précédemment cités sont susceptibles d'apparaître (choix orienté des cas, faible puissance statistique). La confiance dans chacun des énoncés peut être érodée par ces biais. En revanche, comme de nombreux énoncés sont examinés, reflétant une diversité de contextes et donc de variations possibles, le renforcement global de la confiance dans la théorie peut se révéler tout aussi important que dans une étude de test d'hypothèses.

Quelques références bibliographiques pour aller plus loin

Gagnon, Yves-Chantal. 2012. *L'étude de cas comme méthode de recherche*. 2nd ed. Québec: Presses de l'Université du Québec.

Gehman, Joel. et Glaser, Vern L.. et Eisenhardt, Kathleen M.. et Gioia, Denny. et Langley, Ann. et Corley, Kevin G.. 2018. « Finding Theory-Method Fit: A Comparison of Three Qualitative Approaches to Theory Building ». *Journal of Management Inquiry* 27 (3): 284-300. <https://doi.org/10.1177/1056492617706029>.

Ridde, Valéry, éd.. 2021. *Vers une couverture sanitaire universelle en 2030?* Éditions science et bien commun. Québec: Canada: Zenodo. <https://doi.org/10.5281/ZENODO.5166925>.

Stake, Robert E. 1995. *The Art of Case Study Research*. Thousand Oaks, CA: SAGE Publications.

Stiles, William B. 2013. « L'utilisation des études de cas pour l'élaboration de la théorie en psychothérapie ». *Psychothérapies* 33 (1): 29-35. <https://doi.org/10.3917/psys.131.0029>.

Turcotte-Tremblay, Anne-Marie. et Ali Gali-Gali, Idriss. et De Allegri Manuela. et Ridde, Valéry.. 2017. « The Unintended Consequences of Community Verifications for Performance-Based Financing in Burkina Faso ». *Social Science & Medicine* 191 (octobre): 226-36. <https://doi.org/10.1016/j.socscimed.2017.09.007>.

Yin, Robert K. 2010. « Analytic Generalization ». In *Encyclopedia of Case Study Research*, par Albert Mills, Gabrielle Durepos, et Elden Wiebe, 6. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. <https://doi.org/10.4135/9781412957397.n8>.

Yin, Robert K. 2018. *Case study research and applications: design and methods*. Sixth edition. Los Angeles: SAGE.

14. Traçage de processus

ESTELLE RAIMONDO

Résumé

Le traçage de processus est une approche d'évaluation basée sur la théorie. À partir de la formulation d'une théorie du changement relative aux processus (TCP), elle recueille des preuves pour déterminer comment l'intervention s'est déroulée dans un cas unique et si elle a contribué de manière plausible aux changements escomptés. Souvent décrite comme une approche qualitative, le traçage de processus peut en fait s'appuyer sur une diversité de méthodes qualitatives et quantitatives. Particulièrement utile pour évaluer des interventions complexes, cette approche permet de déterminer « dans quelles conditions, comment et pourquoi » une intervention a fonctionné, plutôt que de quantifier son impact.

Mots-clés : Méthodes qualitatives, évaluation basée sur la théorie, théorie du changement, processus, principes causaux, chemins de contribution, preuves, empreintes digitales, raisonnement bayésien

I. En quoi consiste cette approche?

Lorsque les évaluateurs ou évaluatrices effectuent un traçage de processus (TP), ils et elles se comportent un peu comme des « détectives ». En appliquant le traçage de processus, on cherche à expliquer, plutôt qu'à simplement décrire, les processus de changement. Pour le dire simplement, on cherche à retracer comment les activités des personnes et collectifs et leurs motivations s'articulent pour produire un

changement dans les comportements et actions des autres. Le traçage de processus s'apparente également à un « travail de détective » sur le plan empirique, puisqu'il consiste à rassembler un ensemble de preuves (ce que D. Beach appelle des « empreintes digitales ») pour déterminer comment l'intervention s'est déroulée dans un cas particulier et si elle a plausiblement contribué à produire les changements escomptés. En termes un peu plus techniques, le traçage de processus est une approche d'évaluation basée sur la théorie permettant d'étudier comment les interventions ont fonctionné dans des cas réels (voir chapitre séparé sur l'évaluation basée sur la théorie). En tant que tel, le traçage de processus appartient à la famille des méthodes qui cherchent à répondre aux questions « comment, pourquoi et dans quelles circonstances » les programmes et les politiques fonctionnent en étudiant comment ils se déroulent dans le monde réel. Visuellement, le traçage de processus cherche à comprendre ce qui se passe « entre » la flèche qui relie les interventions et les résultats, à partir d'une démarche d'étude de la théorie du changement. Son avantage comparatif par rapport à d'autres méthodes est d'ouvrir complètement la boîte noire des processus de changement.

Le traçage de processus est souvent considéré comme une approche « qualitative » parce qu'il s'appuie souvent sur des preuves qualitatives (provenant d'entretiens, d'observations, de documents, etc.) mais, comme beaucoup d'autres approches d'évaluation basées sur la théorie, il résiste en réalité à une classification simple et peut être plus adéquatement décrit comme « agnostique ». Le traçage de processus peut mobiliser une gamme de méthodes de collecte et d'analyse de données, quantitatives ou qualitatives, en cherchant à rassembler un ensemble de preuves suffisamment solides pour trancher entre la théorie du changement examinée et les explications alternatives. En outre, plus récemment, certains évaluateurs ont formalisé mathématiquement l'utilisation du traçage de processus par l'application du raisonnement bayésien (Befani 2021).

Le traçage des processus se décompose en deux phases principales, et se distingue des autres évaluations basées sur la théorie par quelques caractéristiques uniques que nous mettrons brièvement en évidence.

La première phase du traçage du processus consiste à formuler une théorie du changement relative aux processus (TCP).

Une TCP est une théorie détaillée de la manière dont une intervention a contribué à un résultat d'intérêt. Il s'agit de décortiquer les activités des acteurs et actrices et collectifs qui, ensemble, constituent le fonctionnement interne des programmes (la flèche). Les acteurs et actrices sont les personnes ou les organisations qui font des choses, tandis que les actions sont ce qu'ils et elles font. Pour comprendre pourquoi les actions d'un-e acteur-ric(e) ont conduit d'autres acteurs et actrices à faire des choses, il faut essayer de rendre aussi explicite que possible ce que Cartwright et Hardie (2012) appellent les *principes de causalité*.

Pour ce faire, il s'agit dans un premier temps de réfléchir à la « contribution » qui aurait pu être produite de manière réaliste par une intervention et aux chemins de contribution plausibles reliant l'intervention à ces contributions. Pour cela, on peut s'appuyer sur la littérature en sciences sociales sur le sujet ou sur les archives ouvertes compilant les résultats d'évaluations déjà réalisées (*evidence repositories*), ainsi que sur les documents produits dans le cadre de la mise en œuvre de la politique étudiée. En ce sens, le traçage de processus ne se limite pas aux objectifs déclarés, mais explore plutôt les différentes voies plausibles, intentionnelles ou non, vers les changements escomptés.

Lorsqu'il s'agit de déterminer quelle contribution aurait pu être produite par une intervention, il est également important d'explorer les explications concurrentes, *en dehors* des activités du programme, qui pourraient également expliquer les résultats.

Le niveau de détail d'une TCP peut varier. Une TCP plus détaillée est nécessaire lorsque l'évaluation cherche à produire des connaissances exploitables qui peuvent aider à la mise en œuvre du projet. En revanche, si l'objectif est de comprendre comment un type d'intervention fonctionne dans plusieurs cas, une TCP simplifiée, d'un niveau de précision moyen, peut être suffisant.

La deuxième phase consiste à tester empiriquement la TCP pour comprendre comment elle a réellement fonctionné dans un cas donné.

Le traçage de processus cherche à tester et à affiner sa théorie en observant comment l'intervention a fonctionné dans un cas donné. Dans le traçage de processus, une TCP détaillée sert de base pour tester empiriquement comment une contribution a été effectivement produite. Cela signifie qu'avant de s'engager dans la collecte réelle de données, il faut anticiper le type « d'empreintes digitales » plausibles laissées par le mécanisme de changement et déterminer le type de preuves dont on a besoin ou que l'on souhaiterait voir pour renforcer la confiance dans la théorie. Il existe deux types de preuves utiles : certaines « doivent être trouvées » pour éviter de fragiliser ou d'invalidier la TCP. D'autres sont des preuves qu'on « aimerait trouver » pour renforcer de manière significative la crédibilité de la TCP.

Dans la réflexion sur ces différents éléments de preuve, il s'agit de ratisser large pour englober une variété de différentes « empreintes digitales » potentielles. Dans le cadre du traçage de processus, chaque élément de

preuve individuel ne nous apprend généralement pas grand-chose, mais combiné à d'autres, il peut agir comme une signature unique. Travailler avec des preuves implique donc souvent une forme de bricolage (pour en savoir plus, voir Beach et Pedersen, 2019 : 232-233).

Une fois que la collecte des données a commencé, une évaluation critique des observations et des preuves réunies doit avoir lieu. Le raisonnement bayésien est souvent utilisé comme cadre logique pour évaluer la force (valeur probante) des preuves collectées, soit de manière informelle, de façon analogue à son utilisation dans les enquêtes criminelles (par exemple Beach et Pedersen, 2019), soit de manière plus formelle par l'application du théorème de Bayes et l'estimation des probabilités de trouver ou de ne pas trouver des preuves (voir par exemple Befani et Stedman-Pryce, 2017). Fondamentalement, dans une démarche de traçage de processus, il faut se poser les questions suivantes :

- Si les « empreintes » attendues ne sont pas trouvées, a-t-on eu bien accès à l'ensemble des données empiriques et peut-on être sûr que nos sources ne nous cachaient pas quelque chose?
- Si les « empreintes » attendues sont trouvées, avons-nous interprété correctement ce que nos sources nous ont dit dans ce contexte, et pouvons-nous leur faire confiance?

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

Lorsque le traçage de processus a fait son entrée dans la pratique de l'évaluation, le domaine de l'évaluation d'impact avait été dominé par des approches (quasi-)expérimentales présentant de forts avantages comparatifs dans l'établissement de l'effet de traitement moyen d'interventions relativement simples dont l'effet pouvait être mesuré quantitativement. Cependant, le besoin s'est fait sentir d'élargir la boîte

à outils de l'évaluation d'impact à d'autres approches qui pourraient répondre à différents types de questions causales, et d'étudier des interventions plus complexes et moins accessibles à la quantification et aux comparaisons contrôlées. Le traçage de processus est alors apparu comme une approche utile pour les évaluations qui cherchent à expliquer les processus de changement et qui sont moins concernées par la question de savoir « dans quelle mesure » une intervention a eu un impact sur un résultat souhaité, et plus par la question de « dans quelles conditions, comment et pourquoi » une intervention a fonctionné dans le monde réel.

Le traçage de processus a été utilisé pour évaluer l'impact d'une série d'interventions, mais il présente un avantage comparatif par rapport à d'autres méthodes pour étudier les interventions complexes ou difficiles à saisir, telles que l'influence des connaissances et du travail sur les données, l'impact des campagnes de plaidoyer et de communication ou encore de la concertation sur la prise de décision, etc. Elle fonctionne également bien pour évaluer l'impact des interventions qui visent à modifier des comportements par le biais de mécanismes de sensibilisation et d'incitation.

Le traçage de processus peut être utilisé pour répondre à divers besoins décisionnels, mais il convient particulièrement bien à la gestion adaptative des interventions, lorsqu'on cherche à tester et à affiner les modalités de mise en œuvre d'un programme dans divers contextes. Il peut également être utile d'utiliser le traçage de processus pendant une phase pilote ou de généralisation d'une intervention, pour évaluer si les mécanismes de changement sont déclenchés lorsque les interventions sont reproduites ou étendues. Cette méthode fonctionne généralement bien en tant qu'approche *in itinere* ou *ex post*.

III. Exemples d'utilisation de cette approche dans le domaine du développement

Parmi des exemples d'utilisation du traçage de processus pour l'évaluation des politiques de développement, on peut citer : son utilisation pour évaluer la soutenabilité des politiques d'appui budgétaire (Orth et al. 2017), pour étudier l'impact de campagnes de plaidoyer sur la préservation de la biodiversité (D'Errico, et al. 2017), ou encore pour comprendre la contribution des mécanismes de participation citoyenne dans l'amélioration des services publics en République dominicaine (Raimondo, 2020).

Dans ce dernier exemple, l'évaluation a cherché à répondre à l'intensification des efforts des organismes d'aide pour favoriser la participation des citoyens et citoyennes à la définition des programmes de développement. La Banque mondiale a décidé en 2014 d'intégrer des activités de participation citoyenne dans tous ses projets où des bénéficiaires directs pouvaient être identifié-e-s. En prenant cet engagement politique, la Banque mondiale a affirmé que la participation citoyenne était non seulement positive en principe, mais qu'elle pouvait également améliorer l'efficacité de ses projets. Pour tester cette hypothèse, l'évaluation a porté sur un cas typique d'organisation de la participation citoyenne pour améliorer la fourniture des services de santé et d'éducation pour les ménages pauvres en République dominicaine. Le fait de décortiquer et de tester les mécanismes causaux sous-jacent aux activités de participation citoyenne a certainement permis à l'équipe d'évaluation de mieux comprendre les rouages comportementaux, opérationnels et institutionnels de l'intervention et les conditions dans lesquelles la participation citoyenne pouvait contribuer à l'amélioration de la qualité des services. Sur la base de cette compréhension fine, l'évaluation a formulé des recommandations pratiques concernant la façon dont les réunions avec les citoyen-ne-s devraient être organisées et par qui pour assurer une boucle de rétroaction efficace et l'amélioration

des services. Cependant, il était nécessaire de compléter le traçage de processus par des comparaisons entre cas afin d'améliorer la validité externe des résultats et leur pertinence politique pour l'ensemble du programme, qui a été mis en œuvre dans plusieurs régions.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

La qualité de la mise en œuvre du traçage des processus dépend de la manière dont la théorie et l'empirie sont réunis. Pour parvenir à un traçage de processus présentant une validité interne élevée, il convient de garder à l'esprit les trois critères suivants :

1. un TCP plus désagrégré et plus fin qui capture les épisodes et les mécanismes clés;
2. des preuves très spécifiques trouvées pour chaque partie du TCP;
3. des sources dignes de confiance et un accès complet aux données empiriques.

En revanche, si le TCP est trop simple ou abstrait, si les preuves trouvées ne sont pas spécifiques ou pouvaient valider d'autres explications, ou si les sources sont trop faibles ou ne sont pas dignes de confiance, la validité interne sera faible.

Pour certaines évaluations, il est également important que les leçons tirées du traçage de processus puissent être transposées à d'autres contextes. Le traçage de processus en lui-même n'a pas une grande validité externe, mais en le combinant avec des comparaisons entre cas, il est possible d'explorer si des processus similaires fonctionnent également dans d'autres cas à travers les contextes.

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

Principaux atouts de l'approche lorsqu'elle est bien mise en œuvre :

- Si les trois critères de qualité énoncés ci-dessus sont respectés, l'application du traçage de processus renforce considérablement la capacité à établir un lien de causalité fort entre les interventions et les résultats observés, tout en ayant un fort pouvoir explicatif sur le « comment » et le « pourquoi » des processus de changement.
- Le traçage de processus fournit un plan clair pour organiser et rendre transparent le processus de collecte et d'évaluation des preuves, ainsi que la triangulation des sources. Ce processus va bien au-delà des approches typiques d'étude de cas et d'autres approches basées sur la théorie. Le traçage du processus permet à la théorie du changement de se déployer de manière vivante et rend plus crédibles les revendications d'impact ou de contribution.
- Il est également plus facile de tirer des « leçons pratiques » d'une étude mobilisant le traçage de processus que de nombreux autres types d'approches d'évaluation. En se concentrant sur les explications causales et les liens entre les actions et les changements de comportements, il aide à concevoir comment ces activités devraient être ajustées ou modifiées pour améliorer les résultats.
- Le traçage de processus présente un avantage comparatif par rapport à d'autres méthodes d'évaluation (d'impact) pour évaluer des interventions qui ne se prêtent pas à la quantification ou à l'expérimentation, comme le dialogue politique, la contribution de la recherche, le travail sur les connaissances et les données, les campagnes de plaidoyer et de communication, etc.

Quelques (dé)limitations de l'approche :

- Le traçage de processus ne permet pas de quantifier l'impact moyen d'une intervention sur un résultat d'intérêt et ne doit pas être utilisé pour atteindre cet objectif.
- Bien que le traçage de processus ne requière pas nécessairement une technicité excessive, la courbe d'apprentissage est abrupte pour en maîtriser les ficelles. Il faut notamment se familiariser avec la mise en place des « tests empiriques » pour mesurer la valeur probante (le caractère unique et la fiabilité) des preuves; il s'agit aussi de faire preuve de rigueur dans la reconstruction de la TCP, et de tirer parti de la littérature existante pour théoriser le changement de comportement lié à des actions spécifiques, etc.
- En soi, le traçage de processus a une faible validité externe et doit être associé à une étude de cas croisée, ce qui peut devenir onéreux et prendre du temps.

Quelques références bibliographiques pour aller plus loin

Beach, Derek. et Brun Pedersen, Rasmus. 2019. *Process Tracing Methods*. Ann Arbor: University of Michigan Press.

Befani, Barbara. et Stedman-Bryce, Gavin. 2017. "Process Tracing and Bayesian updating for impact evaluation". *Evaluation*, 23(1): 42–60.

Befani, Barbara. 2021. *Credible Explanations of Development Outcomes: Improving Quality and Rigour with Bayesian Theory-Based Evaluation*. Report 2021:03, Expert Group for Aid Studies (EBA), Sweden.

Cartwright, Nancy. et Hardie, Jeremy. 2012. *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

- D'Errico Stefano. et Befani, Barbara. et Booker, Francesca. et Guiliani, Alessandra. 2017. *Influencing policy change in Uganda: an impact evaluation of the Uganda Poverty and Conservation Learning Group's work*. PCLG Research Report <https://www.iied.org/sites/default/files/pdfs/migrate/G04157.pdf>
- Orth, Magdalena. et Schmitt, Johannes. et Krisch, Franziska. et Oltch, Stefan. 2017. *What we know about the effectiveness of budget support*. Evaluation Synthesis, German Institute for Development Evaluation (DEval), Bonn.
- Raimondo, Estelle. 2020. "Getting Practical with Causal Mechanisms: The Application of Process-Tracing under Real-World Evaluation Constraints." *New Directions for Evaluation*, Fall 2020, 45-58.

15. L'analyse historique comparée

EMANUELE FERRAGINA

Résumé

L'analyse historique comparée combine deux perspectives méthodologiques majeures des sciences sociales, la comparaison (l'étude des similitudes et des différences entre les cas) et l'histoire (l'analyse des processus de changement dans leur dimension temporelle), pour aider à expliquer les phénomènes sociaux à grande échelle sur une variété de sujets. Elle est particulièrement utile pour rendre compte de la définition des politiques publiques (cadre des politiques et changement de politique).

Mots-clés : Méthodes mixtes, méthodes qualitatives, analyse historique, similitudes, différences, histoire, macro, comparaison, point d'inflexion, dépendance au sentier emprunté

I. En quoi consiste cette approche?

L'analyse historique comparée (AHC) est plus une approche qu'une méthode; elle trouve ses origines dans une longue histoire allant d'anciens travaux, par exemple *De la Démocratie en Amérique* (Tocqueville 1960) et *The Protestant Ethic and the Spirit of Capitalism* (Weber 2001), jusqu'à des classiques plus modernes, par exemple *The Social Origins of Dictatorship and Democracy* (Moore 1966) et *States and Social Revolutions* (Skocpol 1979). L'approche historique en sciences sociales propose des explications

de résultats observés à grande échelle sur un large éventail de sujets, tels que les révolutions, l'avènement d'un régime démocratique ou autoritaire, les processus institutionnels de dépendance au sentier emprunté, la continuité et le changement des politiques dans divers domaines. Cette approche présente plusieurs caractéristiques distinctives qui ont favorisé son utilisation intensive dans la recherche en sciences sociales et dans les politiques publiques.

L'AHC explore les similitudes et les différences entre différents cas — rappelant la méthode des accords et des différences de John Stuart Mill — dans le but de dévoiler les mécanismes causaux qui déterminent des résultats spécifiques (voir le chapitre séparé sur les études de cas). Les processus de changement et leur dimension temporelle sont au cœur de la sociologie et de la science politique, et c'est pourquoi l'AHC a contribué à l'identification de l'origine de réformes spécifiques, ou du point de départ de changements institutionnels importants. Les cas analysés sont souvent des États-nations, mais d'autres entités (telles que les régions, les mouvements sociaux et les organisations) ont également été examinées (pour un exemple d'analyse régionale, voir Ferragina 2012; 2013). Cette approche attribue un rôle essentiel à la théorie, et un débat très intéressant a eu lieu dans l'*American Journal of Sociology*, avec un symposium comparant la place assignée à la théorie dans la sociologie historique et la théorie du choix rationnel : « we're no angels : realism, rational choice, and relationality in social science » (voir les contributions à ce débat de Somers 1998; Kiser et Hechter 1998; Goldstone 1998; Calhoun 1998). Le débat a opposé l'utilisation de ces différentes perspectives, en soulignant que l'AHC aide à tester et à générer des théories grâce à une approche macro-configurationnelle, basée sur des cas et orientée dans le temps.

La composante macro concerne les résultats observés à grande échelle, tels que la construction de l'État, les transitions démocratiques, les systèmes d'inégalité, la guerre et la paix. Les chercheurs et chercheuses se concentrent sur les facteurs causaux à grande échelle, y compris les structures politico-économiques (par exemple, le colonialisme) et les

arrangements institutionnels et organisationnels complexes (par exemple, les régimes de politique sociale). Cette approche macro peut également expliquer les événements et les processus de niveau micro qui devraient (ou ne devraient pas) être présents dans des cas particuliers si la théorie macro est correcte. La composante configurationnelle fait référence à la manière dont les chercheurs et chercheuses considèrent comment de multiples facteurs se combinent pour former des ensembles causaux cohérents. Par exemple, on ne peut pas étudier les révolutions sans analyser comment divers événements et processus sous-jacents constituent ces phénomènes sociaux. Même lorsque l'on s'intéresse en AHC à l'étude des effets d'une variable spécifique, on se soucie beaucoup du contexte et des autres causes potentielles.

Contrairement à d'autres techniques couramment utilisées en sciences sociales, l'AHC ne recule pas devant des questions complexes pour lesquelles les données ne sont pas facilement disponibles. La sélection des questions sur la base des données disponibles est en effet l'une des tendances les plus regrettables en sciences sociales. Comme dans la métaphore nietzschéenne, c'est comme si les chercheurs étaient semblables à des personnes ivres qui ne cherchent leurs clés perdues que sous le réverbère. Pour cette raison, l'AHC se concentre sur des questionnements qui se présentent dans le monde réel et utilise des explications basées sur des mécanismes, en suivant des questions du type : pourquoi des cas similaires sur de nombreuses dimensions clés présentent-ils des résultats différents sur une variable dépendante d'intérêt? Ou encore, pourquoi des cas apparemment disparates ont-ils tous le même résultat? De plus, des questionnements peuvent être identifiés dans la réalité empirique lorsque des cas particuliers ne se conforment pas aux attentes de la théorie existante ou de la recherche à grande échelle. L'AHC met l'accent sur le développement d'une compréhension approfondie des cas afin de départager les hypothèses concurrentes.

Sans prétendre à l'exhaustivité, il est important de mentionner ici les outils conceptuels les plus utilisés dans l'AHC, à savoir les points d'inflexion (*critical junctures*), la dépendance au sentier emprunté (*path dependency*) et d'autres dispositifs permettant de saisir le changement graduel. Collier et Collier (1991 : 29) ont défini les points d'inflexion comme des périodes de changement important qui produisent des effets durables. Ces points d'inflexion déstabilisent les modèles institutionnels antérieurs et ouvrent une nouvelle période de dépendance au sentier emprunté. La dépendance au sentier emprunté indique que lorsqu'une nation, ou une autre macro-unité d'analyse, a commencé à se déplacer dans une direction, les coûts pour inverser la trajectoire sont très élevés et cela contribue à une sorte d'inertie qui ne peut être rompue qu'avec un nouveau point d'inflexion (Pierson 2004). En termes simples : l'histoire compte.

Alors que les points d'inflexion et la dépendance au sentier emprunté sont utilisés pour décrire la succession d'un changement radical et de la stabilité, d'autres outils conceptuels décrivent un changement graduel qui peut progressivement produire un changement significatif. Streeck et Thelen (2005) ont distingué cinq catégories pour rendre compte de ces types de changement. Le premier est le déplacement, c'est-à-dire lorsqu'une structure institutionnelle traditionnelle est progressivement discréditée et mise à la marge en faveur de celles qui sont plus aptes à satisfaire les besoins actuels. Ensuite, la superposition renvoie à l'ajout progressif de nouveaux éléments à l'ancienne structure. Cette forme de changement institutionnel est souvent observée dans les politiques sociales, par exemple dans le domaine du marché du travail et de la politique familiale (Daly et Ferragina 2018). Troisièmement, le changement institutionnel peut se produire simplement parce qu'une institution devient obsolète pour répondre à ses objectifs initiaux car elle n'a pas été suffisamment mise à jour au fil du temps : cette forme de changement institutionnel est appelée dérive (Hacker 2004). Une autre forme de changement institutionnel est celle de la conversion, c'est-

à-dire lorsqu'une institution existante est réorientée vers de nouveaux objectifs. Une dernière forme est celle de l'épuisement, qui amène l'institution à une disparition progressive.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

L'AHC peut être utilisée pour comprendre comment mettre en place une étude d'évaluation de politiques publiques, pour identifier les origines de politiques spécifiques, pour mieux comprendre le contexte dans lequel les politiques et les résultats changent, ou encore pour observer une trajectoire institutionnelle sur le long terme. En un mot, une AHC peut aider à situer des évaluations de politiques spécifiques dans un contexte, en illustrant par exemple la concaténation de changements de politiques publiques qui entraînent un changement institutionnel fondamental à long terme (à cet égard, voir l'exemple ci-dessous sur le « néolibéralisme sélectif »). Parmi des ouvrages qui absolvent ces fonctions, citons *Les trois mondes de l'État-providence* (Esping-Andersen 1990), *Development and Crisis of the Welfare State* (Huber et Stephens 2001), *Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment* (Pierson 1994), et *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States* (Skocpol 1992).

III. Démêler l'orientation des réformes de la politique sociale à long terme : le cas du « néolibéralisme sélectif »

L'AHC peut être utilisée pour démêler la manière dont plusieurs réformes peuvent conduire à des résultats spécifiques, en reliant un concept théorique à l'exploration du changement de politique. C'est le cas d'une étude publiée dans *New Political Economy* qui explore comment l'Italie a progressivement libéralisé les politiques de retraite et du marché du travail en différentes étapes (Ferragina et Arrigoni 2021). Si l'on analyse les réformes une par une, on ne peut pas observer correctement la conception globale du processus de libéralisation. Cela signifie qu'une analyse historique pourrait nous permettre de discerner l'ensemble du processus de réforme. L'étude, bien que n'analysant que le cas italien, se base sur la comparaison avec d'autres pays européens à travers la prise en compte du passage de la phase fordiste à la phase néolibérale du capitalisme. Plus précisément, cette recherche illustre le processus italien d'adaptation institutionnelle néolibérale dans les principales réformes de la politique sociale, et suggère que sur trois décennies, ce processus s'est déroulé de manière sélective. Le néolibéralisme sélectif est défini comme une modalité d'adaptation institutionnelle qui est partie des marges pour s'étendre ensuite au reste de la société.

Le néolibéralisme sélectif est le résultat d'un processus de réforme entamé au début des années 1990, lorsque le tournant néolibéral a été amorcé (Ferragina et al. 2022). Le processus de réforme, avec une continuité entre les coalitions de centre-droit et de centre-gauche, a contourné la résistance des syndicats contre une libéralisation globale de la politique sociale, touchant d'abord les groupes sociaux sans ressources de pouvoir suffisantes pour défendre leurs droits sociaux. Cette modalité d'adaptation institutionnelle peut être observée dans les réformes du marché du travail et des pensions.

Grâce au concept de néolibéralisme sélectif, la dualisation initiale des droits sociaux dans le cas italien est interprétée comme une étape intermédiaire vers la libéralisation (pour une discussion, voir Streeck 2009, Emmenegger 2014). Cet argument est étayé par une analyse de la continuité des réformes de la politique sociale et par les enseignements de l'analyse historique comparée. Les idées néolibérales, promues à l'origine par Einaudi dans la première partie du vingtième siècle et maintenues en vie dans les cercles intellectuels de l'après-guerre, sont réapparues comme une rivière souterraine lorsque le contexte international de l'économie politique s'est globalement détourné du keynésianisme. La diffusion des idées néolibérales a influencé les élites technocratiques italiennes à la Banque d'Italie et au Trésor, ainsi que le débat interne des partis socialiste (PSI) et chrétien-démocrate (DC) depuis les années 1980.

L'étude retrace le « démantèlement » du fordisme et le « déploiement » du néolibéralisme et, grâce à cette analyse institutionnelle historique, elle identifie un tournant néolibéral en 1992. Différents courants de la littérature ont souligné l'importance de cette année pour l'Italie – qui peut être considérée comme un tournant sur les plans institutionnel, économique et politique. La notion de « point d'inflexion » est utilisée pour illustrer comment, après 1992, l'équilibre institutionnel a été rompu et a donné lieu à une série de réformes très éloignées du passé. D'un point de vue méthodologique, de tels moments peuvent être qualifiés de « points d'inflexion » parce qu'ils « placent les arrangements institutionnels sur des chemins ou des trajectoires, qui sont ensuite très difficiles à modifier » (Pierson 2004 : 135). Cet outil analytique permet d'identifier une transition du fordisme au néolibéralisme telle qu'elle est dépeinte dans la littérature d'économie politique internationale. Ensuite, le concept de néolibéralisme sélectif aide à interpréter les réformes du marché du travail et des retraites de manière holistique. Cette notion peut être appliquée à d'autres pays et contextes politiques, notamment

lorsqu'une forte résistance des acteurs du veto est mise à mal par un processus de réforme incrémentale qui contribue à une adaptation néolibérale.

IV. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

L'AHC présente des avantages et des inconvénients par rapport à d'autres méthodes et approches. Elle est unique en ce qu'elle aide à aborder les grandes questions et l'analyse des processus politiques, ce qui lui permet de démêler systématiquement des processus de réforme complexes, comme nous l'avons montré avec l'exemple du néolibéralisme sélectif. L'application d'une approche historique permet de considérer avec soin la spécificité des cas, d'observer leur développement à long terme, en proposant au final des généralisations contingentes. Cependant, l'AHC présente également plusieurs limites. L'approche ne propose pas une manière systématique d'aborder certains problèmes comme d'autres méthodes d'analyse. Il est difficile de sélectionner des cas pour tester des théories, et la généralisation, bien que possible, doit être contingente et limitée (à cause du petit N). De plus, cette approche est critiquable d'un point de vue historique, car elle se base souvent sur des sources secondaires plutôt que sur des documents d'archives.

D'autres grandes questions restent ouvertes pour qui voudrait utiliser cette approche à l'avenir. Comment gérer la tension entre structure et agentivité? S'il est essentiel de traiter de ces questions de niveau macro, l'AHC n'offre pas beaucoup d'espace au rôle des individus et s'intéresse principalement au changement structurel. La tension entre la généralisation contingente et le respect des cas analysés soulève aussi des difficultés épistémologiques. Il y a presque soixante ans, Moore (1966 : XIV) a décrit ce problème avec acuité :

Néanmoins, il existe toujours une forte tension entre l'exigence de rendre justice à l'explication d'un cas particulier et la recherche de généralisation, principalement parce qu'il est impossible de connaître l'importance d'un problème particulier tant que l'on n'a pas fini d'examiner tous les problèmes.

Références bibliographiques

- Calhoun, Craig. 1998. Explanation in historical sociology: Narrative, general theory, and historically specific theory. *American journal of sociology*, 104(3): 846-871.
- Collier, Ruth Berins. et Collier, David. 1991. *Shaping the political arena: Critical junctures, the labor movement, and regime dynamics in Latin America*. Princeton: Princeton University Press.
- Daly, Mary. et Ferragina, Emanuele. 2018. Family policy in high-income countries: Five decades of development. *Journal of European Social Policy*, 28(3): 255-270.
- Emmenegger, Patrick. 2014. *The power to dismiss: trade unions and the regulation of job security in Western Europe*. Oxford: Oxford University Press.
- Esping-Andersen, Gosta. 1990. *Three Worlds of Welfare Capitalism*. Cambridge: Polity.
- Ferragina, Emanuele. 2012. *Social capital in Europe: A comparée regional analysis*. Edward Elgar.
- Ferragina, Emanuele. 2013. The socio-economic determinants of social capital and the mediating effect of history: Making Democracy Work revisited. *International Journal of comparée Sociology*, 54(1): 48-73.

- Ferragina, Emanuele. and Arrigoni, Alessandro. 2021. Selective neoliberalism: How Italy went from dualization to liberalisation in labour market and pension reforms. *New Political Economy*, 26(6): 964-984.
- Ferragina, Emanuele. and Arrigoni, Alessandro. and Spreckelsen, Thees. 2022. The rising invisible majority: Bringing society back into political economy. *Review of International Political Economy* 29(1): 114-151.
- Goldstone, Jack. 1998. Initial conditions, general laws, path dependence, and explanation in historical sociology. *American journal of sociology*, 104(3): 829-845.
- Hacker, Jacob. 2004. Privatizing risk without privatizing the welfare state: The hidden politics of social policy retrenchment in the United States. *American Political Science Review*, 98(2): 243-260.
- Huber, Evelyne. and Stephens, John. 2001. *Development and crisis of the welfare state. Parties and policies in global markets*. Chicago: Chicago University Press.
- Kiser, Edgar. and Hechter, Michael. 1998. The debate on historical sociology: Rational choice theory and its critics. *American Journal of Sociology*, 104(3): 785-816.
- Moore, Barrington. jr. 1966. *Social origins of democracy and dictatorship*: Boston: Beacon. *Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press.
- Pierson, Paul. 1994. *Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment*. Cambridge: Cambridge University Press.
- Pierson, Paul. 2004. *Politics in time: history, institutions, and social analysis*. Princeton: Princeton University Press.
- Skocpol, Theda. 1979. *States and social revolutions. A comparative analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

- Skocpol, Theda. 1992. *Protecting soldiers and mothers: The political origins of social policy in the United States*. Cambridge: Harvard University Press.
- Somers, Margaret. 1998. Symposium on Historical Sociology and Rational Choice Theory » We're No Angels »: Realism, Rational Choice, and Relationality in Social Science. *American journal of sociology*, 104(3): 722-784.
- Streeck, Wolfgang. 2009. *Re-forming capitalism: Institutional change in the German political economy*. Oxford: Oxford University Press.
- Streeck, Wolfgang. and Thelen, Kathleen. (2005). *Beyond continuity: Institutional change in advanced political economies*. Oxford: Oxford University Press.
- Tocqueville, Alexis De 1960. *De la démocratie en Amérique*. London: MacMillan & Co Ltd.
- Weber, Max. 2001. *The protestant ethic and the spirit of capitalism*. Chicago: Fritzroy Dearborn Publishers.

Quelques références bibliographiques pour aller plus loin

- Capoccia, Giovanni. et Kelemen, R. Daniel. 2007. **The study of critical junctures: Theory, narrative, and counterfactuals in historical institutionalism.** *World politics*, 59(3): 341-369. Cet article fournit une analyse complète des points d'inflexion. Les points d'inflexion placent les arrangements institutionnels sur des chemins ou des trajectoires, qui sont très difficiles à modifier.

Mahoney, James. et Thelen, Kathleen. (Eds.). 2015. *Advances in Comparative-Historical Analysis*. Cambridge: Cambridge University Press. Cet ouvrage collectif couvre les multiples utilisations de l'analyse historique comparée en science politique. Il comprend des contributions d'auteurs et autrices de premier plan dans le domaine et aborde le vaste programme de l'AHC à travers une analyse des travaux fondamentaux, des outils d'analyse temporelle (tels que la dépendance au sentier et les points d'inflexion), et des développements méthodologiques importants.

Moore, Barrington. Jr. 1966. *Social Origins of Democracy and Dictatorship: Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press. Cet ouvrage fondamental explique les rôles politiques variés joués par la classe supérieure foncière et la paysannerie dans la transformation des sociétés agraires en sociétés industrielles modernes. D'un point de vue méthodologique, Moore souligne la forte tension entre les exigences de rendre justice à l'explication d'un cas particulier et la recherche de généralisation. Un point de départ pour tous ceux qui s'intéressent à l'AHC.

Pierson, Paul. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton: Princeton University Press. L'ouvrage présente une analyse détaillée de l'importance du temps pour comprendre le changement institutionnel et social, apportant un soutien méthodologique à l'affirmation classique selon laquelle l'histoire compte. Pierson suggère d'utiliser l'analyse historique comparée pour aller au-delà d'une vision statique du changement institutionnel.

Skocpol, Theda. 1979. *States and social revolutions: A comparative analysis of France, Russia and China*. Cambridge: Cambridge University Press. Selon Skocpol, les révolutions sociales méritent une attention particulière en raison de leur importance extraordinaire pour l'histoire des nations et de leur modèle distinctif de changement sociopolitique. Ce qui est unique aux révolutions sociales, c'est que les changements fondamentaux dans la structure sociale et politique se

produisent ensemble et se renforcent mutuellement. Pour analyser ces événements historiques importants, Skocpol a procédé à une AHC de la France, de la Russie et de la Chine. Ce livre est une référence pour celles et ceux qui veulent appliquer l'analyse historique comparée à des phénomènes sociaux de grande ampleur.

PARTIE III
MÉTHODES MIXTES ET
APPROCHES
TRANSVERSALES

16. Les méthodes mixtes

PIERRE PLUYE

Résumé

Les méthodes mixtes consistent à intégrer des méthodes qualitatives et quantitatives dans une évaluation ou une recherche. La démarche suppose de réfléchir à cette intégration à toutes les étapes du projet, de la formulation des questions de recherche à l'analyse des données, en passant par la revue de littérature. Les méthodes mixtes permettent un apport descriptif, explicatif ou prédictif supérieur à ceux des méthodes qualitatives ou quantitatives prises séparément.

Mots-clés : Méthodes mixtes, intégration, devis séquentiel exploratoire, devis séquentiel explicatif, devis convergent, revue mixte de la littérature

I. En quoi consistent ces méthodes?

Tout programme peut être évalué en combinant le pouvoir des mots (sons et images) et celui des chiffres (Pluye et Hong 2014). Par exemple, vous pouvez recueillir des histoires auprès des intervenant·e·s et des usagèr·e·s qui illustrent des succès ou des échecs dont on peut tirer des leçons pratiques (ancrées dans l'expérience des parties prenantes) pour améliorer une intervention; de plus, vous pouvez recueillir les statistiques disponibles sur cette intervention, ou planifier leur collecte de manière transversale (par exemple, avec une enquête) ou longitudinale (par exemple, avec une collecte routinière de données insérée dans les

activités quotidiennes). L'intégration des histoires et des statistiques constitue un moyen puissant pour répondre aux défis et questions complexes posés par les politiques publiques.

Dans les sections suivantes, la démarche utilisant les méthodes mixtes est présentée suivant les différentes étapes de la recherche.

Formuler clairement des questions spécifiques

Les méthodes mixtes permettent de répondre à des questions d'évaluation ou de recherche qualitatives et quantitatives interdépendantes (par exemple, séquentielles) ou complémentaires (par exemple, convergentes) sur une politique publique. Par exemple, vous pouvez formuler un objectif général mixte combinant exploration et mesure, puis des questions qualitatives et quantitatives spécifiques (Tableau 1). Toute question doit être formulée clairement. Elle exprime une seule idée (une phrase interrogative). Les questions d'évaluation et de recherche proviennent habituellement des problèmes et des défis rencontrés lors de la création, du développement, de la mise en œuvre (par exemple, l'adaptation au contexte) et de la pérennisation (par exemple, l'ajustement aux changements du contexte) des politiques publiques. Elles sont imposées par la direction ou suggérées par les intervenant·e·s et les usagè·e·s.

Effectuer une revue mixte de la littérature

Toute évaluation ou recherche est guidée par les connaissances existantes. Celles-ci proviennent des experts, de la littérature grise (par exemple, les rapports des organisations publiques identifiables avec Google Scholar ou OpenAlex) et des publications indexées dans les bases

de données bibliographiques comme Cairn, Érudit, Scopus, etc. L'aide d'une documentaliste est inestimable. Commencez par effectuer une revue des revues de littérature publiées sous forme d'articles scientifiques, ou de chapitre de livre ou de thèse. Identifiez les documents les plus pertinents (ceux qui répondent à vos questions) et planifiez une mise à jour des connaissances. Utilisez un logiciel de gestion des documents pour garder une trace du processus et faciliter la rédaction des sections « Introduction » et « Discussion » de votre rapport (par exemple, le logiciel libre Zotero).

Pour mettre à jour les connaissances, les revues mixtes combinent des études quantitatives, qualitatives et/ou mixtes. Elles sont de plus en plus populaires car elles permettent de répondre à des questions qualitatives et quantitatives en tirant profit de la complémentarité des connaissances qualitatives, quantitatives et mixtes. Lorsqu'une politique publique et ses effets sont bien connus, elles permettent d'en fournir une compréhension approfondie et complète dans plusieurs contextes. La grande majorité des revues de littérature ne sont pas systématiques (chères et chronophages), mais les revues mixtes peuvent être systématiques, comme tout autre type de revue : pour en savoir plus sur cette méthode, voir la fiche séparée sur les revues de littérature mixtes.

Choisir un plan (devis) utilisant les méthodes mixtes

L'évaluation et la recherche utilisant les méthodes mixtes s'inspirent généralement de trois devis ou plans de base : séquentiel exploratoire, séquentiel explicatif, et convergent (voir Tableau 2).

Le devis séquentiel exploratoire [QUAL → QUAN] commence avec la collecte et l'analyse de données qualitatives (QUAL). Dans ce devis, les résultats de la phase 1 qualitative informent la collecte et l'analyse des données de la phase 2 quantitative (QUAN). La phase 2 est ainsi fondée sur la perspective des participant·e·s. Ce devis implique d'abord l'exploration

qualitative du phénomène d'intérêt, puis l'utilisation des résultats qualitatifs pour guider l'échantillonnage et la construction de l'outil de collecte de données quantitatives subséquentes (intégration).

Dans le devis séquentiel explicatif [QUAN → QUAL], la collecte et l'analyse des données quantitatives (phase 1) précèdent et informent la collecte des données qualitatives (phase 2). Ce devis implique une évaluation quantitative initiale suivie d'une exploration qualitative de ces résultats, de sorte que les résultats qualitatifs contribuent à l'explication de résultats quantitatifs inattendus ou extrêmes (intégration).

Le devis convergent [QUAN + QUAL] est le plus fréquemment utilisé. Il combine les méthodes qualitatives et quantitatives de manière indépendante et complémentaire. Autrement dit, la collecte et l'analyse des données qualitatives et quantitatives ne dépendent pas l'une de l'autre. Elles sont menées simultanément ou non. En effet, il est rare d'avoir suffisamment de ressources pour tout mener de front. La convergence (intégration) survient au moment de l'interprétation des résultats qualitatifs et quantitatifs. Ce devis implique la collecte de données qualitatives et quantitatives pour répondre à une question similaire formulée de manière qualitative et quantitative.

Collecter et analyser des données

La collecte et l'analyse des données doit tenir compte des sources de données disponibles et des techniques spécifiques, qualitatives ou quantitatives, nécessaires à leur analyse. Certaines procédures peuvent être mixtes, par exemple la technique Delphi (combinant entretiens et questionnaires avec un échantillon de taille moyenne incluant des expert·e·s du monde entier). Étant donné que de nombreuses procédures et techniques d'analyse statistiques et qualitatives peuvent être utilisées, cette fiche est centrée sur l'intégration des méthodes qualitatives et quantitatives.

Stratégies d'intégration

Planifiez toute combinaison pertinente des stratégies pour intégrer les phases (connexion), les résultats (comparaison) et les données (assimilation) qualitatives et quantitatives. En partant d'une revue méthodologique, nous avons identifié trois types d'intégration et neuf stratégies opérationnelles (trois par type d'intégration) pour mener à bien l'intégration des méthodes qualitatives et quantitatives en méthodes mixtes. De plus, nous avons identifié toutes les combinaisons possibles de ces stratégies (Pluye et al. 2018). Ces combinaisons ont été confirmées dans la littérature sur les soins de première ligne, les soins infirmiers, et les sciences de l'éducation, de l'environnement et de l'information. Pour aller plus loin, des techniques spécifiques d'intégration sont décrites dans un manuel (Fetters 2020).

II. En quoi ces méthodes sont-elles utiles pour l'évaluation des politiques publiques?

Les méthodes mixtes sont développées dans plusieurs domaines depuis les années 1970. Elles formalisent des procédures et des techniques pour intégrer les méthodes qualitatives et quantitatives en évaluation et en recherche (Pluye et al. 2019). Elles permettent ainsi d'obtenir une compréhension supérieure à la somme des connaissances obtenues séparément avec des méthodes qualitatives et quantitatives. Par exemple, elles permettent de répondre à la fois à des questions touchant les effets et les coûts des interventions, les processus qui les sous-tendent de même que les expériences et les perspectives des parties prenantes.

III. Un exemple d'utilisation des méthodes mixtes dans le secteur de la santé

Une agence gouvernementale d'Évaluation des Technologies en Santé (ETS) produit et diffuse des recommandations (par exemple des guides sur l'utilisation optimale des médicaments et des normes sur la gestion des services sociaux) à l'échelle nationale via les associations professionnelles, les services sociaux et les services de santé. La direction de l'agence met en œuvre une recherche évaluative qui vise à justifier la pérennisation de cette intervention (imputabilité). Pour chaque recommandation disponible sur le site Internet de l'agence, un questionnaire validé (Granikov et al. 2020) permet aux usagèr·e·s d'en évaluer la pertinence, l'impact cognitif, par exemple en termes d'apprentissage, et l'intention de l'utiliser. En deux ans, plus de 6000 réponses ont été soumises et analysées (statistiques descriptives). De plus, des entretiens sont menés avec 15 usagèr·e·s pour identifier les effets de l'utilisation des recommandations (analyse thématique). L'intégration des statistiques et des thèmes permet d'estimer les retombées de l'intervention (utilisation et effets), et d'ajouter des types d'effets attendus dans le questionnaire.

IV. Quels sont les critères permettant de juger de la qualité de ces méthodes?

Les méthodes mixtes doivent satisfaire à trois conditions nécessaires ou caractéristiques essentielles : (a) au moins une méthode qualitative et une méthode quantitative sont intégrées; (b) chaque méthode est utilisée de façon rigoureuse par rapport aux critères généralement admis dans la méthodologie ou la tradition de recherche invoquée; et (c) l'intégration des méthodes est effectuée au minimum au moyen de questions d'évaluation ou de recherche, d'un devis (plan) et d'une stratégie

d'intégration des résultats ou des données qualitatives et quantitatives. Quelques outils permettent d'évaluer la qualité des méthodes mixtes en appliquant ces principes. Leur liste est mise à jour sur le site catevaluation.ca. L'outil validé le plus populaire est disponible gratuitement sur Internet (Hong et al. 2018) : il comprend une grille de vérification, un manuel d'utilisation et des réponses aux questions les plus fréquentes (mixedmethodsappraisaltoolpublic.pbworks.com).

Par ailleurs, de nombreux guides et manuels facilitent la rédaction d'un rapport d'évaluation ou d'une publication scientifique utilisant les méthodes mixtes (Creswell et Plano Clark 2018). Leur liste est mise à jour sur le site equator-network.org. Les recommandations GRAMMS (« Good Reporting of a Mixed Methods Study ») listent six éléments essentiels à inclure dans un document qui rapporte l'utilisation des méthodes mixtes (O'Cathain, Murphy, et Nicholl 2008) : (a) justifier l'usage de ces méthodes en lien avec les questions de recherche; (b) indiquer le plan (séquentiel ou convergent) d'usage des méthodes mixtes; (c) détailler les méthodes qualitatives et quantitatives mobilisées; (d) préciser quand, comment et qui a procédé à l'intégration des méthodes mobilisées; (e) présenter les limites de ces méthodes; et (f) indiquer quels ont été les apports des différentes méthodes, ainsi que l'apport complémentaire de leur intégration.

V. Quels sont les atouts et les limites de ces méthodes par rapport à d'autres?

Les avantages des méthodes mixtes résident dans la synergie entre méthodes qualitatives et quantitatives. L'intégration de ces méthodes donne une valeur ajoutée aux méthodes prises séparément (Fetters et Freshwater 2015). En revanche, les méthodes mixtes entraînent un travail supplémentaire pour collecter et analyser à la fois des mots (sons et images) et des statistiques, et pour intégrer les données et résultats

qualitatifs et quantitatifs. Leur mobilisation peut donc prendre plus de temps qu’une seule méthode, et nécessite une équipe multidisciplinaire avec au moins un-e expert-e pour chacune des méthodes sélectionnées. Finalement, elles requièrent plus d’espace dans une publication.

Tableau 1. Questions qualitatives et quantitatives

Question	Description et exemples
	Centrée sur un seul phénomène.
Qualitative	<p>Quoi, pourquoi ou comment. Par exemple, « Qu'est-ce que représente le retour aux études du point de vue des gestionnaires des hôpitaux universitaires de Toulouse ».</p> <p>Verbe exploratoire (par exemple, comprendre, découvrir, décrire, explorer, identifier).</p> <p>Indication de : Politique étudiée; Contexte dans lequel elle est étudiée; Type de données (par exemple, l'expérience de vie); Interprétation des données.</p>
Quantitative descriptive	<p>Étude dite d'incidence ou de prévalence</p> <p>Par exemple, « Combien de gestionnaires des services de santé sont retourné-e-s aux études en 2022 dans les pays francophones? » (données recueillies par les universités; pays, type de service, ancienneté, genre, et ressource dédié au retour).</p>
Quantitative inférentielle	<p>Par exemple, « Quelle est l'importance (et la probabilité) de l'influence des facteurs familiaux et sociaux sur ce retour aux études? »</p> <p>Indication de : Population étudiée et échantillonnage; Intervention ou exposition à une politique; Groupe de contrôle ou de comparaison; Effets en fonction de la durée de l'intervention ou de l'exposition; Hypothèse (verbe suggérant une forme de causalité ou de lien théorique ou logique comme affecter, associer, causer, influencer); Paramètres mesurés.</p>

Tableau 2. Trois devis de base utilisés en méthodes mixtes : Exemples

Devis Exemple de la politique/intervention « Bourse de reprise des études »

Séquentiel exploratoire

[QUAL → QUAN]

- Phase 1 : Entretiens menés avec des gestionnaires avant l'intervention.
- Connexion des phases : Résultats utilisé pour construire l'intervention (politique incitative) et son évaluation.
- Phase 2 : Recueil des statistiques avant/après l'intervention.

Séquentiel explicatif

[QUAN → QUAL]

- Phase 1 : Recueil des statistiques avant/après l'intervention.
- Connexion des phases : Identification des boursier(ère)s qui n'ont pas complété les études prévues (A), ou ont décliné une bourse (B).
- Phase 2 : Entretiens avec les gestionnaires A (barrières aux études?) et B (insuffisance de la bourse?).

Convergent

[QUAN + QUAL]

- Entretiens avec un échantillon raisonné de gestionnaires (raisons pour lesquelles l'intervention est suffisante ou insuffisante?)
 - Simultanément, une enquête mesure l'importance et la probabilité de l'influence des facteurs associés avec la reprise des études auprès d'un échantillon représentatif des gestionnaires cibles de cette politique.
 - Comparaison des résultats qualitatifs et quantitatifs : Par exemple, la politique a les effets escomptés pour un coût raisonnable (efficience), mais peut être bonifiée en tenant compte des raisons pour lesquelles certain-e-s gestionnaires clés jugent la bourse insuffisante.
-

Quelques références bibliographiques pour aller plus loin

Creswell, John. et Vicki, Plano Clark. 2018. Designing and conducting mixed methods research. 3rd éd. Thousand Oaks: SAGE.

- Fetters, Michael. 2020. *The mixed methods research workbook: Activities for designing, implementing, and publishing projects*. Thousand Oaks: SAGE.
- Fetters, Michael. et Freshwater, Dawn. 2015. « The 1+ 1= 3 integration challenge ». *Journal of Mixed Methods Research* 9 (2): 115-17.
- Granikov, Vera. et Grad, Roland. et El Sherif, Reem. et Shulha, Michael. et Chaput, Genevieve. et Doray, Genevieve. et Lagarde, François. et Rochette, Annie. et Tang, David Li. et Pluye, Pierre. 2020. « The Information Assessment Method: Over 15 years of research evaluating the value of health information. » *Education for Information* 36 (1): 7-18.
- Hong, Quan Nha. et Fàbregues, Sergi. et Bartlett, Gillian. et Boardman, Felicity. et Cargo, Margaret. et Dagenais, Pierre. et Gagnon, Marie-Pierre. et Griffiths, Frances. et Nicolau, Belinda. et O'Cathain, Alicia. 2018. « The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. » *Education for information* 34 (4): 285-91.
- O'Cathain, Alicia. et Murphy, Elizabeth. et Nicholl, Jon. 2008. « The quality of mixed methods studies in health services research. » *Journal of Health Services Research and Policy* 13 (2): 92-98.
- Pluye, Pierre. et Bengoechea, Enrique García. et Granikov, Vera. et Kaur, Navdeep. et Tang, David Li. 2018. « Tout un monde de possibilités en méthodes mixtes : revue des combinaisons des stratégies utilisées pour intégrer les phases, résultats et données qualitatifs et quantitatifs en méthodes mixtes. » Dans *Oser les défis des méthodes mixtes en sciences sociales et sciences de la santé*, sous la direction de Bujold, Mathieu. et Hong, Quan Nha. et Ridde, Valéry. et Bourque, Claude Julie. et Dogba, Maman Joyce. et Vedel, Isabelle. et Pluye, Pierre. 28-48. Montréal: Association francophone pour le savoir.

- Pluye, Pierre. et Bengoechea, Enrique García. et Tang, David Li. et Granikov, Vera. 2019. « La pratique de l'intégration en méthodes mixtes. » Dans *Évaluation des interventions de santé mondiale: méthodes avancées*, sous la direction de Ridde Valéry et Christian Dagenais, 213-38. Québec: Éditions science et bien commun.
- Pluye, Pierre. et Hong, Quan Nha. 2014. « Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. » *Annual Review of Public Health* 35: 29-45.

17. Les revues de littérature systématiques mixtes

QUAN NHA HONG

Résumé

Les revues de littérature systématiques mixtes consistent à procéder à un examen systématique et à dresser un bilan des travaux disponibles (et notamment des évaluations déjà réalisées) sur un sujet donné, en intégrant des travaux mobilisant des méthodes qualitatives, quantitatives et mixtes. Ce type de revue de littérature permet notamment une meilleure compréhension des interventions et phénomènes complexes. En englobant une diversité de questions évaluatives, elle est particulièrement utile pour éclairer la prise de décision publique.

Mots-clés : Méthodes mixtes, études qualitatives, études quantitatives, revue systématique mixte, revue de la littérature

I. En quoi consiste cette méthode?

La démarche de revue de littérature consiste à résumer, combiner, analyser, commenter ou critiquer des écrits portant sur un sujet donné. Les revues de littérature systématiques mixtes ont comme particularité d'inclure une variété de types d'études pour mieux comprendre des phénomènes complexes : des études quantitatives (par exemple, des études qui mesurent les effets d'une intervention ou l'ampleur d'un problème), des études qualitatives (par exemple, des études qui portent

sur l'expérience des personnes) et des études utilisant des méthodes mixtes (c'est-à-dire, études qui utilisent des méthodes quantitatives et qualitatives).

Les revues systématiques mixtes font partie de la grande famille des revues systématiques de la littérature, c'est-à-dire un type de revue de littérature qui vise à répondre à une question de recherche en suivant une démarche préalablement définie pour le repérage, la sélection, l'évaluation et la synthèse des études pertinentes. Il s'agit d'un type considéré parmi les plus rigoureux puisqu'il permet de minimiser des erreurs et des biais qui peuvent survenir durant le processus de la revue. Ces revues utilisent donc des méthodes explicites et les rapportent de manière transparente afin qu'elles puissent être reproduites. Outre la revue systématique, il existe une variété de types de revues qui utilisent une approche systématique tels que la revue de la portée (*scoping review*), la revue rapide (*rapid review*) et la revue de revues systématiques (*overview of reviews*).

De manière générale, la mise en œuvre d'une revue systématique mixte suit huit étapes :

1. Formuler une ou des questions de recherche — la formulation des questions peut être guidée par une exploration sommaire de la littérature existante sur le sujet d'intérêt;
2. Définir des critères d'éligibilité (inclusion et exclusion) pour la sélection des articles;
3. Identifier des sources documentaires pour s'assurer d'avoir une recherche exhaustive telles que chercher dans des bases de données bibliographiques (par exemple, PubMed, Health Policy Reference Center, International Political Science Abstracts, Europa World, Worldwide Political Sciences Abstracts, Web of Science, JSTOR, SocINDEX), consulter la table de matière de périodiques scientifiques

sur le sujet d'intérêt, chercher dans des sites web sur le sujet, consulter la liste de références des articles retenus ou les articles qui ont été cités, et contacter des expertes et experts;

4. Élaborer une stratégie de recherche documentaire avec la collaboration d'un ou d'une bibliothécaire spécialisé-e qui maîtrise les techniques de recherche documentaire dans les bases de données et les autres sources;
5. Sélectionner des documents pertinents en suivant deux étapes soit la sélection à partir des titres et résumés des documents puis la sélection à partir des textes intégraux;
6. Évaluer la qualité des documents retenus à l'aide d'outils pour évaluer la qualité des études;
7. Extraire les données des documents retenus à l'aide d'un formulaire qui précise toutes les données à extraire; et
8. Synthétiser les données extraites, c'est-à-dire analyser les données récoltées dans la revue pour en faire un tout cohérent pour répondre à la ou aux questions de recherche. Dans le cadre des revues systématiques mixtes, la synthèse doit aussi considérer comment les données qualitatives et quantitatives seront intégrées. De manière générale, deux principaux devis (ou protocoles) de synthèse peuvent être utilisés pour synthétiser les données qualitatives et quantitatives : a) devis de synthèse convergent dans lequel les données sont synthétisées de manière simultanée (par exemple, les études quantitatives et les études qualitatives sont synthétisées séparément avec différentes méthodes de synthèse puis les résultats des deux synthèses sont comparés) et b) devis de synthèse séquentiel qui implique au moins deux phases de synthèses dépendantes (par exemple, les études qualitatives sont d'abord synthétisées puis les résultats de cette synthèse informent la synthèse des études quantitatives).

Pour de plus amples informations sur chacune de ces étapes et sur les devis de synthèse, vous pouvez consulter les références à la fin de ce document ainsi que ce site web : <http://toolkit4mixedstudiesreviews.pbworks.com>.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

Les revues systématiques mixtes sont pertinentes pour l'évaluation des politiques publiques, car elles permettent une compréhension plus approfondie des phénomènes et des interventions complexes. Les phénomènes complexes sont souvent caractérisés par une multiplicité d'acteurs impliqués ainsi qu'une diversité de modèles d'intervention et de facteurs qui influencent leur succès. Divers types d'études peuvent être utilisés pour évaluer ces phénomènes complexes. Ainsi, la synthèse de ces études complémentaires permettra d'acquérir une compréhension plus globale de l'état des connaissances sur un phénomène complexe.

La revue systématique mixte permet de fournir un portrait plus large et complet de la littérature sur un sujet donné. Elle permet aussi de combiner des questions complémentaires telles que : Quelle est l'efficacité d'une politique? Pourquoi cette politique est-elle efficace ou pas? Comment la politique fonctionne-t-elle? Quels sont les facteurs qui entravent ou facilitent l'implantation de cette politique? Pour répondre à ces questions, il est nécessaire d'inclure des études quantitatives pour répondre aux questions sur l'efficacité des politiques et des études qualitatives qui porteront sur les questions sur le pourquoi et sur le comment. Les réponses à ces questions complémentaires peuvent favoriser une meilleure prise de décision chez les personnes décideuses politiques, gestionnaires et praticiennes.

D'autres avantages des revues systématiques mixtes ont été soulignés comme de permettre une meilleure compréhension des résultats obtenus dans les études quantitatives à partir des études qualitatives (ou vice versa), de considérer une diversité de perspectives (par exemple, perspectives des personnes décideuses et utilisatrices), de corroborer des connaissances obtenues, et de renforcer la crédibilité et la validité des conclusions.

III. Un exemple d'utilisation de cette méthode : la lutte contre le tabagisme chez les jeunes

Une revue systématique mixte a été effectuée par des chercheurs et chercheuses de l'EPPI-Centre (*Evidence for Policy and Practice Information and Co-ordinating Centre*) au Royaume-Uni pour informer le développement de politiques visant à réduire le taux de tabagisme chez les jeunes (Sutcliffe, Twamley, Hinds et al., 2011). Dans leur revue, ils et elles se sont intéressé-e-s à trois principales questions de recherche :

- a) Quels sont les modes d'accès aux produits du tabac vendus au détail et non au détail les plus utilisés par les jeunes âgé-e-s de 11 à 18 ans et les modes varient-ils en fonction de facteurs tels que l'âge et le sexe?
- b) Quelle est la perception des jeunes sur l'accès aux produits du tabac et quels sont, selon les jeunes, les obstacles et les facteurs facilitant l'accès aux produits du tabac? et c) Quels types d'interventions visant à limiter l'accès aux produits du tabac aux jeunes en dehors du commerce de détail ont été évalués et comment ces interventions abordent-elles les obstacles et les facilitateurs identifiés comme importants par les jeunes au Royaume-Uni?

Pour cette revue systématique mixte, une recherche documentaire dans plus d'une centaine de sources d'information a été effectuée (par exemple, bases de données, sites web, recherche par citations). Cette recherche a permis d'identifier six études qualitatives, sept sondages et seize études

d'intervention. Deux personnes ont été impliquées de manière indépendante dans la sélection des études, l'évaluation de leur qualité, l'extraction et la synthèse des données. Les synthèses pour chaque type d'étude inclus (sondages, études qualitatives, études d'intervention) ont été effectuées séparément. Pour l'intégration, les résultats de ces synthèses ont par la suite été comparés de deux manières : a) évaluer le niveau de concordance entre les résultats des sondages et ceux des études qualitatives concernant les sources des produits du tabac des jeunes et leurs modes d'accès par sexe, âge et statut tabagique (occasionnel ou régulier) et b) évaluer dans quelle mesure les interventions ont permis d'aborder les obstacles et les facilitateurs identifiés par les jeunes dans les études qualitatives.

À la lumière des résultats de cette revue, trois principales implications pour le développement de politiques pour réduire le taux de tabagisme chez les jeunes ont été formulées. Premièrement, les études ont révélé qu'il était facile d'accéder aux produits du tabac par des personnes du réseau amical et des pairs (accès social) dans les écoles. Les jeunes ont décrit cet accès social comme étant omniprésent, organisé et visible. Ainsi, le développement d'une réglementation plus stricte dans les écoles pour réduire l'accès social devrait être exploré. Deuxièmement, les résultats des études indiquent que la mise en œuvre d'une réglementation dans les commerces de détail était variable. Il est donc nécessaire d'explorer les raisons de la mise en œuvre inégale et d'identifier les moyens pour permettre un meilleur contrôle de la mise en œuvre de la réglementation dans ces commerces de détail. Troisièmement, les résultats de cette revue suggèrent la nécessité de s'attaquer aux achats de produits du tabac par une tierce personne adulte comme des personnes du réseau familial et amical et des personnes extérieures.

Dans cette revue systématique mixte, l'utilisation de données provenant de différents types d'études a permis d'identifier différents modes d'accès utilisés, de mieux comprendre les expériences et les opinions des jeunes sur l'accès aux produits du tabac et d'explorer des pistes d'intervention potentielles. Aussi, la nature mixte de la revue, qui combine des données

d'enquêtes, des recherches sur le point de vue des jeunes au Royaume-Uni et des interventions portant sur l'accès aux produits du tabac en dehors du commerce de détail, a permis de fournir des données contextualisées pour l'élaboration de politiques.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Pour juger la qualité des revues systématiques, il est important de comprendre les sources d'erreurs et de biais potentiels qui peuvent influencer les résultats obtenus. Nous en présenterons quatre : biais de repérage, biais de divulgation, biais de sélection et biais d'interprétation.

Le biais de repérage survient lorsque les études pertinentes sur le sujet d'intérêt ne sont pas repérées. Ce biais est relié à la recherche documentaire et aux sources de repérage. Afin de faire une recherche exhaustive pour repérer l'ensemble des études pertinentes pour la question de recherche, il est important de diversifier les sources documentaires, d'utiliser différentes bases de données et de développer des stratégies de recherche documentaire rigoureuse avec la collaboration de bibliothécaires spécialisés.

Le biais de divulgation, dont le plus connu est le biais de publication, survient lorsque la nature, la direction ou la force des résultats d'une étude influencent sa publication. Par exemple, il a été démontré que les études ayant des effets positifs ont plus de chance d'être publiées et être publiées plus rapidement dans des périodiques scientifiques que celles qui démontrent des résultats négatifs. Ceci peut entraîner une surreprésentation des études démontrant des effets positifs et peut affecter les conclusions de la revue systématique. Pour minimiser ce biais, il est recommandé de diversifier les sources de données et les types

de documents à inclure comme des rapports scientifiques provenant de centres de recherche et des thèses et mémoires rédigés dans le cadre des études supérieures universitaires.

Un biais de sélection survient lorsque la sélection des études est arbitraire ou influencée par des motivations ou convictions particulières. Par exemple, une chercheuse ou un chercheur pourrait croire qu'une intervention est importante et décider d'inclure seulement des études qui ont démontré que l'intervention est efficace. Ceci biaisera donc les résultats de la revue. Pour minimiser ce biais, il est important de définir des critères d'éligibilité clairs préalablement à la sélection et d'impliquer deux personnes lors de la sélection.

Le biais d'interprétation est relié à la personne qui évalue et son interprétation des études. Ce biais peut être minimisé en impliquant au moins deux personnes dans l'extraction des données, l'évaluation de la qualité des études et la synthèse des données. De plus, dans une revue systématique mixte, il est recommandé d'avoir une équipe avec des expertises complémentaires en méthodologies de recherche qualitative et quantitative afin de faciliter la synthèse et le jugement sur la qualité des études.

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

Une revue systématique mixte permet de répondre à des questions diverses, de tirer profit de la complémentarité des données quantitatives et qualitatives, et d'avoir une compréhension approfondie et complète d'un phénomène complexe. Aussi, le fait d'utiliser une méthodologie rigoureuse et explicite permet de minimiser des erreurs et des biais potentiels qui pourraient influencer la validité des résultats d'une revue. Toutefois, divers défis peuvent survenir lors de son opérationnalisation.

Un défi important est le temps et les ressources requis. La durée d'une revue systématique peut varier entre 6 à 24 mois. Divers facteurs peuvent influencer sa durée tels que les questions de recherche à traiter, le nombre de personnes impliquées, le nombre de documents à analyser, et la ou les méthodes de synthèse à utiliser. Aussi, le fait d'inclure une diversité de types d'études dans une revue systématique mixte augmente le volume de document à identifier, trier, extraire et analyser. Il est ainsi important de s'assurer de la disponibilité des ressources et de bien justifier le choix d'effectuer une revue systématique mixte.

Les questions qui pourront être étudiées dans une revue systématique dépendent de la littérature disponible. Par exemple, dans le cadre d'une revue systématique mixte, une équipe de recherche pourrait être intéressée à identifier des études sur les effets d'une intervention et d'autres sur la perspective des utilisateurs sur l'intervention. Imaginons toutefois que la recherche documentaire ne permette d'identifier que des études sur les effets et aucune sur la perspective des utilisateurs sur l'intervention d'intérêt. Dans cet exemple, la revue n'est pas mixte puisque seulement un type d'études est synthétisé. Afin d'orienter les questions de recherche spécifiques qui pourraient être abordées sur un sujet d'intérêt dans le cadre d'une revue systématique mixte, il peut être utile de mener une exploration préalable et sommaire de la littérature existante.

Un autre défi concerne l'intégration des données, c'est-à-dire comment les composantes qualitatives et quantitatives sont combinées. Il s'agit d'une caractéristique clé de la revue systématique mixte qui permet que l'ensemble des résultats des divers types d'études retenus soient intégrés afin de fournir une compréhension plus approfondie sur le sujet d'intérêt et émettre des recommandations qui reflètent l'ensemble de la littérature couverte. Une revue qui n'a pas d'intégration pourrait être considérée comme incluant plusieurs revues indépendantes plutôt qu'une revue mixte. Il est donc essentiel de bien décrire la façon dont les données sont intégrées et de bien refléter la valeur ajoutée de cette intégration et de ses limites.

Quelques références bibliographiques pour aller plus loin

Heyvaert, Mieke. et Hannes, Karin. et Onghena, Patrick. 2016. « Using Mixed Methods Research Synthesis For Literature Reviews: The Mixed Methods Research Synthesis Approach. » Thousand Oaks, CA: SAGE Publications.

Hong, Quan Nha. et Pluye, Pierre. et Bujold, Mathieu. et Wassef, Maggy. 2018. « Les défis des revues systématiques mixtes : devis de synthèse convergents et séquentiels. » Dans *Cahier scientifique Acfas #117 – Oser les défis des méthodes mixtes en sciences sociales et sciences de la santé*, sous la direction de Mathieu Bujold, Quan Nha Hong, Valéry Ridde, Claude Julie Bourque, Maman Joyce Dogba, Isabelle Vedel et Pierre Pluye, 49-63. Montréal : Association francophone pour le savoir (Acfas). <https://www.acfas.ca/publications/magazine/2018/03/cahier-scientifique-117-methodes-mixtes>.

Hong, Quan Nha. et Rees, Rebecca. et Sutcliffe, Katy. et Thomas, James. 2020. « Variations of mixed methods reviews approaches: A case study. » *Research Synthesis Methods* 11 (6): 795-811. <https://doi.org/10.1002/jrsm.1437>.

Hong, Quan Nha. et Turcotte-Tremblay, Anne-Marie. et Pluye, Pierre. 2019. « Les revues systématiques mixtes – Un exemple à propos du financement basé sur les résultats. » Dans *Évaluation des interventions de santé mondiale. Méthodes avancées*, sous la direction de Valéry Ridde et Christian Dagenais, 157-202. Québec : Éditions science et bien commun. <https://scienceetbiencommun.pressbooks.pub/evalsantemondiale/chapter/revuesmixtes/>.

Lizarondo, Lucylynn. et Stern, Cindy. et Carrier, Judith. et Godfrey, Christina. et Rieger, Kendra. et Salmond, Susan. et Apostolo, Joao. et Kirkpatrick, Pamela. et Loveday, Heather. 2020. « Chapter 8: Mixed

methods systematic reviews. » Dans JBI Manual for Evidence Synthesis, sous la direction d'E. Aromataris et Z. Munn. Adelaide : JBI. <https://jbi-global-wiki.refined.site/space/MANUAL>.

Sutcliffe, Katy. et Twamley, Katherine. et Hinds, Kate. et O'Mara, Allison. et Thomas, James et Brunton, Ginny. 2011. « Young people's access to tobacco: A mixed-method systematic review. » EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London (London). <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3301>.

18. Les comparaisons de niveau macro

EMANUELE FERRAGINA

Résumé

Les comparaisons de niveau macro sont une approche qui exploite les variations et les similitudes entre de grandes unités d'analyse macrosociales (par exemple, des États, des régions, des provinces) pour étudier différents phénomènes sociaux. Les études peuvent être entreprises à différentes échelles et à des fins diverses, par exemple pour décrire les différences de niveau macro entre différents États, ou pour évaluer l'influence d'une structure différente de l'État-providence sur les résultats individuels (tels que les niveaux de chômage, l'espérance de vie, etc.).

Mots-clés : Méthodes mixtes, unités macro-sociales, variation, similitudes, État-providence

I. En quoi consiste cette approche?

Toute enquête scientifique est intrinsèquement comparative, ce qui est clairement observable lorsque l'on considère les méthodes les plus courantes dans les sciences sociales. Pour donner quelques exemples : les expérimentations sont comparatives parce qu'elles ont besoin d'un groupe de contrôle pour mesurer l'effet d'une intervention par comparaison avec son absence; les analyses de régression contrôlent l'effet de plusieurs variables en comparant leur effet sur une série de cas.

Par conséquent, si toutes les méthodes de recherche sont comparatives au sens large, en sciences sociales, l'idée d'enquête comparative fait souvent strictement référence à des recherches impliquant l'utilisation de grandes unités d'analyse macro-sociales (Ragin, 2014). La recherche dans ce sens est comparative lorsqu'elle exploite la variation ou la similarité entre des unités macro-sociales d'analyse, par exemple un état, une région, une province¹. Cela peut ensuite donner lieu à des études basées sur différents niveaux et échelles, mais toutes incluent l'utilisation de macro-unités d'analyse. L'objectif de ces macro-comparaisons est de comprendre la complexité causale et de décrire la relation entre les macro-unités d'analyse et les micro-unités d'analyse et entre les macro-unités d'analyse entre elles. La littérature en fournit différents exemples, par exemple les comparaisons entre différents modèles de sécurité sociale, ou l'évaluation de l'impact d'une configuration spécifique de la politique familiale sur les taux d'emploi et de fécondité des femmes. L'analyse des unités macrosociales est une « catégorie méta-théorique » qui distingue fondamentalement les scientifiques adoptant cette approche, qui utilisent « des unités macrosociales dans les énoncés explicatifs (et descriptifs) » (Ragin, 2014 : 5). En effet, la grande majorité des scientifiques travaillant dans le domaine (y compris l'auteur de ce

1. Ainsi, par exemple, la modélisation à plusieurs niveaux est incluse dans cette définition. Cependant, la définition de la recherche comparative de Ragin, fondée sur des unités d'analyse macro-sociales, n'est pas universellement acceptée. D'autres chercheurs ou chercheuses ont proposé différentes limites pour délimiter le domaine de la recherche comparative. D'une part, celles et ceux qui sont plus orienté-e-s vers l'utilisation de techniques quantitatives et multivariées ont défini la méthode comparative en considérant simplement les études qui incluent des données comparatives provenant de différentes sociétés (Andreski, 1965; Armer, 1973) ou les travaux basés sur l'analyse multiniveau (Rokkan, 1966; Przeworski et Teune, 1970). D'autre part, les chercheurs et chercheuses plus versé-e-s dans l'analyse qualitative/historique tels que Moore (1966) et Skocpol (1979) ont tendance à faire la distinction entre les méthodes comparatives basées sur les cas et celles orientées vers les variables (la lignée remonte bien sûr aux pères fondateurs de la sociologie et de la science politique, par exemple Tocqueville, Durkheim et Weber). Nous suggérons que ces points de vue sont trop restrictifs pour nos objectifs, et pour cette raison, avec Ragin (2014), nous définissons la méthode comparative et les macro comparaisons sur la base de leur objectif principal.

chapitre!), ne définissent souvent pas la nature et le rôle des unités macro-sociales, mais les utilisent plutôt implicitement comme des unités d'analyse « d'observation » et/ou « explicatives » (Ragin, 2014 : 8).

Par conséquent, l'utilisation de macro-comparaisons est davantage un mode de pensée qu'une méthode *stricto sensu*. Les macro-comparaisons peuvent être établies à l'aide de différentes techniques au niveau quantitatif, qualitatif et historique, par exemple les statistiques descriptives, les études de cas et l'analyse historique comparée (AHC), l'analyse qualitative comparée (ou comparative qualitative analysis, QCA) (voir fiches dédiées sur ces trois méthodes), les ensembles flous (*fuzzy sets*), les techniques de régression, les modèles d'équations structurelles (MES) et les analyses factorielles, ainsi que l'analyse en grappes (*cluster analysis*). D'autres techniques utilisées moins fréquemment sont les modèles de référence diagonaux, l'analyse séquentielle, la construction d'échelles, l'analyse thématique, l'appariement par score de propension (PSM), l'appariement optimal, l'alpha de Krippendorff (KA) et l'analyse de l'historique des événements (pour un examen systématique des méthodes utilisées dans la recherche macro comparative, voir Ferragina et Deeming 2022). Cela signifie que les macro-comparaisons ne sont pas limitées à des techniques spécifiques, mais qu'elles doivent plutôt être considérées comme une structuration de la « pensée sur la pensée » (Sartori 1970) afin d'accroître l'inférence (les conclusions plus larges qui peuvent être tirées) que nous tirons de l'étude de cas spécifiques.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

Les comparaisons de niveau macro sont extrêmement utiles pour l'évaluation des politiques publiques, tant *ex ante* qu'*ex post*. En particulier, elles peuvent aider à contextualiser les résultats issus d'études de cas ou d'expérimentations spécifiques. Pour faire avancer le débat sur

la relation entre des politiques spécifiques et leurs effets, il est essentiel que les comparaisons de niveau macro et les études de cas nationales puissent apprendre les unes des autres (Ferragina 2020). Les études de cas nationales — par exemple l'évaluation d'une politique spécifique au sein d'un pays — souffrent souvent d'un manque de validité externe (la capacité de généraliser les conclusions au-delà du cas étudié). À l'inverse, lorsqu'ils et elles utilisent des expérimentations, les chercheurs et chercheuses sont en mesure de tester l'effet de réformes graduelles, mais pas l'effet global d'un domaine de politique publique sur un résultat spécifique. Ainsi, par exemple, dans le domaine de la politique familiale, les comparaisons de niveau macro peuvent aider à démêler l'impact de l'effet conjoint de politiques familiales explicites différentes (c'est-à-dire la garde d'enfants, les congés parentaux et les prestations familiales) sur l'emploi des femmes dans les différents pays, tandis que l'expérimentation peut permettre de démêler l'effet spécifique d'une augmentation du nombre de structures de garde d'enfants sur l'élasticité de l'emploi des femmes dans un cas spécifique. Pour cette raison, nous avons besoin de plus d'études qui adossent systématiquement l'étude de mesures politiques au contexte dans lequel elles sont mises en œuvre. En ce sens, les comparaisons de niveau macro peuvent non seulement offrir des perspectives intéressantes sur les effets de différentes politiques au niveau transnational ou transrégional, mais aussi permettre d'évaluer de manière critique les résultats d'évaluations spécifiques. En outre, d'un point de vue explicatif, l'existence de données macro comparatives consolidées peut aider à interpréter les résultats d'études menées au niveau national. C'est le cas de l'un des travaux macro comparatifs les plus célèbres jamais publiés, à savoir *Les trois mondes de l'État-providence* de Gøsta Esping-Andersen (1990).

III. Les trois mondes de l'État-providence : un exemple célèbre d'apport de la méthode comparative à l'évaluation de politiques publiques

L'ouvrage *Les trois mondes de l'État-providence* s'inscrit dans une longue tradition académique en sociologie et en science politique, ancrée dans le raisonnement déductif² et l'utilisation d'idéaux-types³. Comme l'a souligné Max Weber (1904 : 87), « la construction d'un système de propositions abstraites et donc purement formelles [...] est le seul moyen d'analyser et de maîtriser intellectuellement la complexité de la vie sociale ». Dans cette veine, Esping-Andersen (1990) a construit la typologie des régimes de protection sociale en reconnaissant l'importance et le pouvoir idéologiques des trois mouvements politiques dominants du long du 20^e siècle en Europe occidentale et en Amérique du Nord, à savoir la social-démocratie, la démocratie chrétienne (conservatisme) et le libéralisme.

L'État providence social-démocrate idéal-typique repose sur le principe de l'universalisme, en accordant l'accès aux prestations et aux services sur la base de la citoyenneté. Un tel État providence est censé offrir un degré d'autonomie relativement élevé, limitant la dépendance à l'égard de la famille et du marché. Afin d'atteindre l'autonomie, les États-providence sociaux-démocrates se caractérisent par un niveau élevé de démarchandisation⁴ et un faible degré de stratification. Les politiques

2. Le raisonnement déductif est une forme de pensée logique qui part d'une idée générale pour arriver à une conclusion spécifique. Il s'agit d'une pensée descendante qui va du général au spécifique.

3. Un idéal-type est une construction analytique dérivée de la réalité observable, mais qui ne s'y conforme pas en détail en raison d'une simplification délibérée. Il est « idéal » car il est utilisé pour se rapprocher de la réalité en sélectionnant et en accentuant certains éléments.

4. La démarchandisation désigne le degré auquel les individus ou les familles peuvent maintenir un niveau de vie socialement acceptable indépendamment de la participation au marché (tel que défini par Esping-Andersen dans *Les trois mondes de l'État-providence*).

sociales sont perçues comme une « politique contre le marché » (Esping-Andersen, 1985). Les États-providence chrétiens-démocrates sont fondés sur le principe de subsidiarité et la prédominance des régimes d'assurance sociale, offrant un niveau moyen de démarchandisation et un degré élevé de stratification sociale. Le régime libéral repose sur la notion de domination du marché et des prestations privées; idéalement, l'État n'intervient que pour atténuer la pauvreté et répondre aux besoins fondamentaux, essentiellement sous condition de ressources. Par conséquent, le potentiel de démarchandisation des prestations publiques est faible et la stratification sociale élevée. Toutefois, ces modèles ne sont pas purs et, dans chaque cas national réel, différentes caractéristiques sont mélangées. En ce sens, Esping-Andersen montre clairement comment le dispositif comparatif est un moyen de classer et de comprendre les différences et les groupes de pays, mais doit être considéré avec prudence :

Si les États-providence peuvent être groupés, il faut reconnaître qu'il n'existe aucun cas tout à fait pur. Les pays scandinaves peuvent être à dominance social-démocrate, mais ils ne sont pas dépourvus d'éléments libéraux. Les régimes libéraux ne sont pas non plus des types purs. Le système de Sécurité sociale américain est un système de redistribution obligatoire et loin d'être totalement actuariel. Au moins dans sa première formulation, le New Deal était aussi social-démocrate que l'est la social-démocratie suédoise contemporaine. Et les régimes conservateurs européens ont incorporé des impulsions libérales et sociales-démocrates. Au fil des décennies, ils sont devenus moins corporatistes et moins autoritaires. (Esping-Andersen, 1990 : 73)

Diverses contributions ont confirmé sa typologie, tandis que d'autres l'ont contestée et élargie, d'un point de vue théorique et méthodologique (voir Ferragina et Seeleib-Kaiser 2011; Ferragina et Filetti 2022 pour une discussion). Cependant, malgré ce long débat et d'importantes controverses dans la littérature, on ne peut nier le rôle fondamental

que cet ouvrage a joué dans la structuration et la compréhension d'un segment important des politiques publiques, à savoir les politiques sociales. En particulier, il offre une représentation souple de l'utilité des comparaisons macro et le cadre développé par Esping-Andersen a été utilisé comme point de départ pour des milliers d'études (au 26 octobre 2022, le livre a été cité 44 086 fois!).

En ce qui concerne l'évaluation des politiques publiques, les travaux d'Esping-Andersen ont été utilisés :

- Pour sélectionner différents cas d'étude pour une analyse. La sélection d'au moins un cas social-démocrate, un cas chrétien-démocrate et un cas libéral permet de tirer plus d'enseignements de l'étude de quelques pays.
- En tant qu'outil heuristique pour interpréter les effets des différentes politiques dans les différents pays.
- Pour comprendre et décrire les différentes trajectoires des pays dans le temps.
- Pour contextualiser les résultats obtenus en comparant différents pays.

IV. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

Les comparaisons de niveau macro sont utilisées pour tester des hypothèses, déduire des liens de causalité, illustrer et comprendre en profondeur des modèles spécifiques et interpréter le changement social. Elles permettent un plus grand pouvoir d'interprétation par rapport aux études de cas individuels. Cela implique un fort pouvoir heuristique. Ce n'est pas un hasard si des ouvrages très cités comme *Les trois mondes de l'État-providence* ont fourni aux chercheurs et chercheuses en politique publique des informations importantes sur un grand nombre de pays

développés, qui restent valables plus de 30 ans après la publication de l'ouvrage d'Esping-Andersen. Les comparaisons de niveau macro permettent de prêter attention au contexte et aux effets potentiels que ce contexte pourrait exercer sur des résultats spécifiques. Cependant, le fait d'examiner « la forêt » plutôt que « les arbres » impose d'une part des coûts élevés (en termes d'expertise sur des cas multiples), et d'autre part une simplification de l'analyse pour permettre les comparaisons entre différentes unités d'analyse de niveau macro. Cela peut engendrer plusieurs problèmes, tels que la classification erronée (la création de pseudo-classes qui simplifient de manière incorrecte l'univers des cas analysés) et « l'étirement conceptuel », c'est-à-dire l'application erronée de théories et de concepts à des cas autres que ceux qui ont été analysés.

Les chercheurs et chercheuses ont souvent tendance à inclure un grand nombre de pays dans leur comparaison en élargissant les catégories qu'ils et elles ont développées sur la base de connaissances directes acquises à travers de quelques cas. Cependant, cet élargissement peut s'avérer problématique à plusieurs égards. D'une part, il est utile d'avoir plus de pays afin de mieux tester une série d'hypothèses, mais d'autre part, avec moins de cas, on peut être plus précis dans la définition des concepts. Ce compromis n'est pas toujours pris en compte dans les sciences sociales contemporaines, avec des comparaisons qui finissent par trop élargir les concepts. Par conséquent, les concepts et les idées tirés des comparaisons de niveau macro doivent être utilisés avec un grain de sel. En tant qu'approche plus que méthode, les comparaisons de niveau macro permettent une approche critique des sciences sociales et ont historiquement soulevé des questions importantes sur les résultats obtenus. En conclusion, elles sont une épée à double tranchant : elles peuvent informer de manière significative l'évaluation des politiques publiques, mais elles doivent également être considérées avec prudence.

Références bibliographiques

- Andreski, Stanislav 1965. *The Uses of Comparative Sociology*. Berkeley: University of California Press.
- Armer, Michael. 1973. « Methodological problems and possibilities in comparative research ». In: Armer, Michael. et Grinmshaw, Qllen. (eds), *Comparative Social Research*. New York: Wiley, 49–79.
- Esping-Andersen, Gosta. 1985. *Politics against markets*. Princeton: Princeton University Press.
- Esping-Andersen, Gosta. 1990. *Les trois mondes de l'État-providence*. Princeton: Princeton University Press.
- Ferragina, Emanuele. 2020. Family policy and women's employment outcomes in 45 high-income countries: A systematic qualitative review of 238 comparative and national studies. *Social Policy & Administration* 54(7), 1016-1066.
- Ferragina, Emanuele. et Deeming, Christopher. 2022. « Methodologies for comparative social policy analysis ». In: Yerkes Mara A. and Nelson Keneth. and Nieuwenhuis, Rense (eds), *Changing European Societies: The Role for Social Policy Research*. Cheltenham: Edward Elgar, 218–235.
- Ferragina, Emanuele. et Filetti, Federico Danilo. 2022. Labour market protection across space and time: a revised typology and a taxonomy of countries' trajectories of change. *Journal of European Social Policy* 32(2): 148–165.
- Ferragina, Emanuele. et Seeleib-Kaiser, Martin. 2011. Welfare regime debate: past, present, futures? *Policy & Politics* 39(4): 583–611.
- Moore, Barrington. Jr. 1966. *Social Origins of Dictatorship and Democracy*. London: Penguin.

- Przeworski, Adam. et Teune, Henry. 1970. *The Logic of Comparative Social Enquiry*. New York: Wiley.
- Ragin, Charles. 2014. *The Comparative Method. Moving beyond Qualitative and Quantitative Strategies*. Oakland: University of California Press.
- Rokkan, Stein. 1966. « Comparative cross-national research ». In: Merritt, Richard. and Rokkan, Stein. (eds), *Comparing Nations*. New Haven: Yale University Press. 3–26.
- Sartori, Giovanni. 1970. Concept misformation in comparative politics. *American Political Science Review* 64(4): 1033–1053.
- Skocpol, Theda. 1979. *States and social revolutions*. Cambridge: Cambridge University Press.
- Skocpol, Theda. 1992. *Protecting soldiers and mothers: The political origins of social policy in the United States*. Harvard: Harvard University Press.
- Weber, Max. 1904. *On the methodology of the social sciences*. Glencoe: The Free Press.

Quelques références bibliographiques pour aller plus loin

- Ferragina, Emanuele. et Deeming, Christopher. (à paraître). Comparative mainstreaming? Mapping the uses of the comparative method in social policy, sociology and political science since the 1970s. *Journal of European Social Policy*.** Une analyse de 50 ans de recherche comparative basée sur une base de données comprenant des milliers d'articles comparatifs issus des meilleures revues de sociologie, de sciences politiques et de sociologie. L'analyse quantitative des

principales tendances dans l'utilisation de la méthode comparative est complétée par une analyse qualitative des articles les plus cités dans le domaine comparatif.

Kohn, Melvin. L. 1987. *Cross-national research as an analytic strategy: American Sociological Association, 1987 presidential address. American sociological review*, 52(6), 713-731. Ce discours présidentiel de l'American Sociological Association suggère que la recherche comparative transnationale est un outil essentiel pour générer, tester et développer la théorie sociologique. La méthode comparative est coûteuse et difficile à appliquer, et elle peut également générer certains problèmes d'interprétation. Cependant, malgré ses limites, elle constitue un outil fondamental de la recherche en sciences sociales.

Lijphart, Arend. 1971. *Comparative politics and the comparative method. American Political Science Review*, 65(3), 682-693. L'article propose une analyse systématique de la méthode comparative. Il met l'accent à la fois sur les limites de la méthode et sur les façons dont, malgré ces limites, elle peut être utilisée de manière optimale. Lijphart se concentre sur le rôle des études de cas (sous leurs différentes formes) comme principal moyen d'entreprendre des macro-comparaisons. Dans l'article, il oppose les comparaisons basées sur des cas aux méthodes expérimentales et statistiques.

Przeworski, Adam. et Teune, Henry. 1970. *The Logic of Comparative Social Inquiry*. New York: John Wiley and Sons. Dans cet ouvrage novateur, les auteurs ont proposé des idées et des points de vue sur la recherche comparative qui ont profondément marqué la recherche en sciences politiques. Le livre se concentre principalement sur l'analyse quantitative. Une lecture indispensable pour tous les étudiants intéressés par la méthode comparative et ce qu'on peut en faire.

Ragin, Charles. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press. Ce livre offre des perspectives considérables pour

la compréhension et l'utilisation de l'analyse comparative. Rédigé à l'origine pour présenter l'utilité de l'analyse comparative qualitative (ACQ) par rapport aux techniques qualitatives et quantitatives, il fournit également des raisons théoriques et de fond pour l'utilisation de la méthode comparative et des macro comparaisons dans le domaine des politiques publiques.

19. L'analyse qualitative comparée

VALÉRIE PATTYN

Résumé

L'analyse qualitative comparée (AQC ou QCA pour *qualitative comparative analysis*) est une méthode mixte qui traduit des données qualitatives en un format numérique afin d'analyser systématiquement quelles configurations de facteurs produisent un résultat donné. L'AQC repose en effet sur une conception configurationnelle de la causalité, selon laquelle les résultats découlent de combinaisons de conditions. Elle est très utile pour l'évaluation d'impact *ex post*, plus précisément pour comprendre pourquoi une même politique peut entraîner certains changements dans certaines circonstances et pas dans d'autres.

Mots-clés : Configurations, combinaisons de conditions, complexité causale, identification systématique de modèles de cas croisés, équi-finalité, causalité conjoncturelle, causalité asymétrique

I. En quoi consiste cette méthode?

Pourquoi une même politique entraîne-t-elle certains changements dans certaines circonstances et pas dans d'autres? Prenons l'exemple d'un programme de subventions destiné à aider les entreprises à dispenser une formation interne sur les compétences managériales. Comment se fait-il qu'une telle formation soit efficace pour certains employés et pas pour d'autres? Ou, autrement dit, dans quelles conditions l'efficacité du

transfert de la formation est-elle assurée ou non? L'analyse qualitative comparée (AQC) est une méthode permettant de répondre à une telle question.

L'AQC part du principe que des configurations — c'est-à-dire des combinaisons de conditions — sont nécessaires et/ou suffisantes pour obtenir un résultat donné. Les conditions peuvent être conçues comme des variables causales, des déterminants ou des facteurs (Rihoux et Ragin, 2009 : xix). Un résultat, dans un contexte d'évaluation, est généralement un effet politique intentionnel ou non intentionnel bien défini qui peut être présent ou absent. Dans l'exemple ci-dessus, le résultat est l'occurrence ou la non-occurrence de « l'efficacité du transfert de la formation ».

Contrairement à d'autres méthodes basées sur des études de cas (voir le chapitre séparé sur les études de cas), l'AQC permet de comparer systématiquement des informations basées sur des cas, et permet donc une généralisation modeste. En même temps, contrairement aux méthodes statistiques, elle nous permet de conserver des informations contextuelles riches et une certaine complexité. En raison de ce double potentiel, la méthode est souvent décrite comme une passerelle entre les méthodes qualitatives et quantitatives. La méthode a été développée à l'origine pour les équipes de recherche confrontées à un nombre intermédiaire de cas (entre 10 et 50), mais elle est de plus en plus souvent appliquée dans des contextes comportant un grand nombre de cas (voir Thomann, et Maggetti 2020).

Il est important de noter que l'AQC n'est pas seulement une technique analytique, mais qu'elle s'accompagne également d'une approche spécifique de la causalité, appelée causalité conjoncturelle multiple, qui est très compatible avec les hypothèses qui sous-tendent l'évaluation réaliste (voir le chapitre séparé sur l'évaluation réaliste). En particulier, cette approche implique que :

- Les effets des politiques publiques sont souvent le résultat de combinaisons de conditions plutôt que le résultat d'une seule condition (« causalité conjoncturelle »).
- Différentes configurations possibles peuvent conduire aux mêmes effets ou résultats observés : c'est ce que l'AQC appelle « l'équifinalité ».
- La causalité est comprise de manière asymétrique : si, dans un cas donné, une certaine condition est pertinente pour le résultat, son absence n'entraîne pas nécessairement l'absence du résultat.

L'AQC appartient à la famille des méthodes de la théorie des ensembles (*set theory*). Un cas peut faire partie d'un ou plusieurs ensembles. Les ensembles articulent les caractéristiques que certains cas peuvent avoir en commun. Dans notre exemple, un employé ou une employée qui participe à une formation en entreprise peut faire partie de l'ensemble des cas des « employé·e·s autonomes dans leur travail et leur prise de décision » et/ou de l'ensemble des « employé·e·s qui ont reçu le soutien de leurs responsables hiérarchiques pour suivre la formation ». En identifiant la mesure dans laquelle un cas fait partie d'un certain ensemble et en le comparant systématiquement à d'autres cas présentant des variations dans l'occurrence d'un certain résultat (c'est-à-dire l'efficacité du transfert de la formation), on peut découvrir quels (combinaisons de) facteurs sont nécessaires et/ou suffisants pour ce résultat :

- Une (combinaison de) condition-s jugée-s nécessaire-s implique qu'elle sera toujours présente/absente chaque fois que le résultat est présent/absent. Ou pour le dire en termes de théorie des ensembles, X est une condition nécessaire pour Y, si Y est un sous-ensemble de X ($X \leftarrow Y$). Par exemple, si nous constatons que toutes les formations en entreprise menant à l'efficacité du transfert de formation ont été dispensées par des instructeurs ou instructrices ayant une grande expérience de l'enseignement, cette dernière peut être qualifiée de condition nécessaire.

- Pour qu'une condition (ou une combinaison de conditions, c'est-à-dire une configuration) soit suffisante, le résultat doit apparaître chaque fois que la condition est présente. En théorie des ensembles, une condition (X) est qualifiée de suffisante si elle constitue un sous-ensemble de l'issue ($X \rightarrow Y$). Par exemple, une formation suivie par un·e employé·e ayant une grande autonomie dans son travail et fortement motivé·e peut constituer un chemin suffisant pour l'efficacité du transfert de la formation.

Les cas peuvent prendre différentes formes en AQC. Dans le cadre d'une évaluation, les cas sont généralement des contextes dans lesquels une intervention a été appliquée. Dans l'exemple mentionné précédemment, les cas concernent les employé·e·s qui ont suivi une formation subventionnée. Les cas peuvent également être des organisations ou des entreprises, ou être situés au niveau macro (c'est-à-dire des pays).

Comment alors comparer systématiquement de tels cas? Pour cela, on peut recourir à différentes techniques d'AQC. Dans l'AQC *crisp set* (AQCcs), la version originale de l'AQC, les conditions et les résultats doivent être traduits en termes binaires, 1 ou 0. C'est ce qu'on appelle la calibration. Les conditions ou les résultats auxquels est attribué un score de 1 doivent être lus comme présents (ou élevés, ou importants), tandis que ceux dont le score est de 0 sont considérés comme absents (ou faibles, ou petits). Les scores binaires expriment des différences qualitatives de nature. Dans la variante de l'AQC basée sur un ensemble flou (*fuzzy set*, AQCfs), les cas peuvent avoir une appartenance partielle à un ensemble et un score compris entre 0 et 1, ce qui tient compte du fait que de nombreux phénomènes sociaux sont dichotomiques « en principe » mais que les manifestations empiriques de ces phénomènes dans la pratique diffèrent souvent en degré (Schneider et Wagemann 2012, 14).

Quelle que soit la technique utilisée, le cycle de recherche de l'AQC passe par des étapes similaires :

Premièrement, les données étant calibrées, on peut construire une matrice de données qui présente essentiellement les données observées empiriquement sous la forme d'une liste de configurations.

Deuxièmement, la matrice de données calibrée peut, dans une étape ultérieure, être transformée en une table dite de vérité (*truth table*), qui énumère toutes les configurations possibles menant à un résultat particulier. Comme une seule configuration peut correspondre à plusieurs cas empiriques, la table de vérité résume donc le tableau des données empiriques. Le nombre total de configurations théoriquement possibles dans la table de vérité est déterminé par le nombre de conditions incluses dans la recherche. Il s'agit de trouver un bon équilibre entre le nombre de cas et de conditions. Les configurations non couvertes par les observations empiriques peuvent être considérées comme des restes logiques, c'est-à-dire qu'elles sont logiquement possibles, mais non observées. L'AQC offre l'opportunité intéressante d'inclure des hypothèses plausibles sur le résultat de (ou d'une sélection de) restes logiques pour tirer des inférences plus parcimonieuses (Schneider, et Wagemann 2012).

Troisièmement, la table de vérité ouvre la voie au moment analytique de l'AQC, appelé minimisation booléenne. Dans ce processus, on peut s'appuyer sur différents logiciels. La minimisation repose sur l'hypothèse que si deux combinaisons ne diffèrent que sur une seule condition, mais présentent le même résultat, cette condition particulière est redondante. Elle peut donc être éliminée pour obtenir une représentation plus simple du cas (ou du groupe de cas). En appliquant cette règle de manière itérative à toutes les paires de combinaisons possibles jusqu'à ce qu'aucune autre simplification ne soit possible, on obtient une série de chemins suffisants vers le résultat. Le type de résultats (c'est-à-dire les formules de solution) résultant généralement de l'analyse AQC seront des expressions sur « la (combinaison de) conditions qui sont nécessaires et/ou suffisantes pour l'occurrence ou la non-occurrence d'un résultat particulier ».

Quatrièmement, et c'est le point le plus important, une étude AQC ne s'arrête pas après l'application du logiciel. Il est essentiel que le chercheur ou la chercheuse explique ensuite le lien de causalité de manière narrative (Schneider et Wagemann 2010), en revenant sur les cas individuels et en reliant les résultats à des connaissances théoriques et conceptuelles plus larges. Cette méthode a en commun sa nature itérative : on peut faire des allers-retours entre l'analyse préliminaire des données et l'ensemble des données ou la théorie du changement. Ce processus est également un moyen utile pour apprendre à connaître les cas plus en profondeur.

II. En quoi cette méthode est-elle utile pour l'évaluation des politiques publiques?

L'AQC peut être utilisée à des fins explicatives, ce qui permet de tester les théories du programme (ou théories du changement) de manière systématique, ou à des fins exploratoires pour développer des théories à partir de connaissances basées sur des cas. Comme le résultat (les effets) doit être connu avant l'évaluation, la méthode ne peut en principe être appliquée que pour des évaluations *ex post* et *in itinere* dans lesquelles une intervention a entraîné une variation du succès.

De par son objectif, l'AQC convient principalement aux évaluations axées sur l'apprentissage plutôt que sur la reddition de comptes. En particulier, on dit souvent qu'elle peut contribuer à l'apprentissage en double et triple boucle : non seulement elle met en lumière les conditions dans lesquelles les interventions politiques fonctionnent, mais elle offre également la possibilité d'impliquer activement les parties prenantes dans le processus de sélection des résultats, des conditions et dans le calibrage de celles-ci. En conséquence, les parties prenantes et les commanditaires peuvent mieux comprendre ce que signifie un « changement réussi » dans le contexte de l'intervention, et ce qui fait la différence.

III. Un exemple d'utilisation de cette méthode dans la politique de formation

Nous avons déjà fait allusion à une évaluation de l'efficacité de la formation aux compétences non techniques (telles que les compétences managériales). Cette évaluation a été menée dans des entreprises flamandes (belges), à la demande de l'agence flamande du Fonds social européen (FSE), qui a également subventionné la formation. Bien que des recherches contrefactuelles antérieures aient démontré l'impact positif des subventions à la formation, elles ont également révélé qu'il n'y a pas toujours de transfert de ce qui est appris dans l'environnement de travail. Cette observation a constitué la principale raison d'être de l'évaluation et a incité le commanditaire à se concentrer non plus sur la question de savoir si les formations subventionnées fonctionnent, mais sur les conditions dans lesquelles les programmes de formation fonctionnent. L'étude a porté sur 50 cas, dont 15 cas réussis dans lesquels les compétences sociales ont été transférées et 35 cas d'échecs, dans lesquels l'efficacité du transfert de formation n'était pas atteinte au moment de l'évaluation.

Sur la base de la littérature pertinente en sciences de l'éducation, 8 conditions ont été identifiées comme ayant un pouvoir explicatif potentiel et incluses dans notre modèle AQC : (1) soutien des pairs; (2) soutien du responsable hiérarchique; (3) sentiment d'urgence; (4) prévention des rechutes et fixation d'objectifs; et les conditions contextuelles suivantes; (5) éléments identiques; (6) programme de formation en tant que méthode d'apprentissage active; (7) autonomie et (8) charge de travail équilibrée. Parmi ces conditions, aucune ne s'est avérée nécessaire à la réussite du transfert de la formation. Cependant, nous avons identifié plusieurs voies consistant en des combinaisons de conditions qui étaient suffisantes pour réussir : chaque fois que ces voies étaient présentes, le contenu de la formation était retenu et appliqué avec succès sur le lieu de travail. Le tableau 1 ci-dessous présente les huit parcours identifiés.

L'analyse AQC s'est avérée utile pour savoir quelles conditions surveiller dans les futurs programmes subventionnés. L'évaluation était une composante d'une évaluation multiméthode : elle a été suivie d'un traçage de processus (voir le chapitre séparé sur le traçage de processus) qui s'est concentré sur une sélection de cas pour identifier les mécanismes par lesquels les programmes de formation font une différence dans l'environnement de travail. L'analyse AQC a également permis d'identifier systématiquement les cas pour lesquels une analyse plus approfondie au sein du cas était la plus pertinente.

Cas	Soutien des pairs	Soutien du superviseur	Sens de l'urgence	Prévention des rechutes et fixation d'objectifs	Éléments identiques	Le programme de formation comme méthode d'apprentissage actif	Autonomie	Charge de travail équilibrée
J3; V2								-
B2; K2							-	
M1; D1		-						
N2; B3			-					
W1								
T1								
S2								
T2								

Tableau 1 : Les voies d'un transfert de formation réussi Note : Blanc : l'état est absent; Gris : l'état est présent; “-“ non inclus dans le parcours. Source : Álamos-Concha, Prischilla. Cambré, Bart. Foubert, Josephine. Pattyn Valérie. Rihoux, Benoit. et Schalembier Benjamin. 2020. Impactevaluatie ESF-interventie Opleidingen in bedrijven. What drives training transfer effectiveness and how does this transfer work? Commissioned by Departement Werk en Sociale Economie. Vlaamse Overheid: p. 11. Full report: [https://www.vlaanderen.be/publicaties/impactevaluatie-esf-interventie-opleidingen-in-bedrijven-what-drives-training-tr](https://www.vlaanderen.be/publicaties/impactevaluatie-esf-interventie-opleidingen-in-bedrijven-what-drives-training-transfer-effectiveness-and-how-does-this-transfer-work)

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette méthode?

Plusieurs *checklists* circulent avec des aperçus de ce qu'impliquent les bonnes pratiques en matière d'AQC (Schneider, et Wagemann 2010; Befani 2016 : 183-185), et cela dépasserait le cadre de cette fiche méthodologique d'élaborer sur tous les critères de qualité. Il est impératif que l'AQC, en tant que technique d'analyse des données, soit appliquée de manière cohérente avec « l'esprit » de l'AQC en tant qu'approche de recherche, ce qui implique que l'AQC ne doit pas être réduite à un « processus à boutons » mécaniste. En outre, il importe d'être transparent·e sur tous les choix effectués dans le processus de recherche, et idéalement de recourir à des tests de robustesse pour toutes les décisions prises. Ce dernier point est particulièrement important, étant donné la forte sensibilité de la méthode au cas par cas.

L'analyse AQC générera différents paramètres d'ajustement qui aideront à évaluer les analyses de nécessité et de suffisance. En termes simples, la cohérence décrit la mesure dans laquelle une relation empirique entre une (combinaison de) condition(s) et le résultat se rapproche de la nécessité et/ou de la suffisance en théorie des ensembles. La couverture décrit l'importance empirique ou la pertinence d'une combinaison de conditions. Pour les conditions nécessaires, la cohérence est généralement fixée à un niveau très élevé, à savoir 0,9, alors que pour les conditions suffisantes, des valeurs de cohérence plus faibles (par exemple 0,75) sont relativement courantes. Les valeurs de couverture doivent généralement être de 0,60 ou plus. Il est toutefois important de noter que les seuils de ce qui est considéré comme « bon » peuvent varier en fonction de la conception et de l'objectif de la recherche (Schneider et Wagemann 2010).

V. Quels sont les atouts et les limites de cette méthode par rapport à d'autres?

L'AQC présente l'avantage unique de tenir compte de la complexité causale, tout en permettant une généralisation modeste par l'identification systématique de schémas croisés. Les procédures rigoureuses sur lesquelles elle s'appuie rendent également les résultats parfaitement reproductibles. Un autre avantage est qu'elle ne nécessite pas un grand nombre de cas pour être appliquée.

Toutefois, à proprement parler, l'AQC ne permet de mettre en évidence que des « associations » entre une condition et un résultat. L'interprétation causale réelle est du ressort des personnes qui mènent l'évaluation. Une limitation similaire s'applique à l'élément temps. Bien que l'on travaille actuellement sur différentes manières d'inclure le « temps » dans une analyse AQC (voir Verweij et Vis, 2021), le type de résultats obtenus est de nature statique plutôt que dynamique.

Pour ces raisons, il est conseillé de combiner l'AQC avec d'autres méthodes intra-cas qui ont la capacité d'ouvrir la boîte noire causale. En particulier, la combinaison de l'AQC et du traçage de processus est de plus en plus utilisée à cette fin. L'évaluation mentionnée dans cette fiche méthodologique en est un exemple.

Quelques références bibliographiques pour aller plus loin

Befani, Barbara. 2016. « Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA). » *Sztokholm: Expertgruppen för biståndsanalys (the Expert Group for Aid*

Studies). Pobrane: <http://eba.se/en/pathways-to-change-evaluating-development-interventions-with-qualitative-comparative-analysis-qca>

Rihoux, Benoît. et Ragin, Charles C.. 2008. Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques. Sage Publications.

Schneider, Carsten. et Wagemann, Claudius. 2010. « Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. » *Comparative sociology* 9, no.3: 397-418.

Schneider, Carsten. et Wagemann, Claudius. 2012. Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis. Cambridge University Press.

Thomann, Eva. et Maggetti, Martino. 2020. « Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools. » *Sociological Methods & Research* 49, no.2: 356-386.

Verweij, Stefan. et Vis, Barbara. 2021. « Three strategies to track configurations over time with Qualitative Comparative Analysis. » *European Political Science Review* 13, no.1: 95-111.

Une excellente source est également <http://www.compasss.org>, qui comprend une vaste bibliographie, un aperçu des logiciels, des tutoriels et des directives sur l'AQC.

20. L'évaluation basée sur la théorie

AGATHE DEVAUX-SPATARAKIS

Résumé

L'évaluation basée sur la théorie s'est développée en réponse aux limites des approches expérimentales et quasi-expérimentales, qui ne permettent pas de saisir les mécanismes par lesquels une intervention produit ses impacts. Cette approche consiste à ouvrir la « boîte noire » de l'action publique en décomposant les différentes étapes de la chaîne causale liant l'intervention à ses résultats finaux. Les hypothèses ainsi formulées sur les mécanismes en jeu peuvent ensuite être testées empiriquement.

Mots-clés : Méthodes qualitatives, théorie, chaîne causale, logique d'intervention, diagramme logique d'impact, logigramme, théorie du changement, chemin d'impact

I. En quoi consiste cette approche?

Formalisée dans les années 1970 et 1980 aux États-Unis, l'évaluation basée sur la théorie (EBT) n'est pas une méthode, mais plutôt une logique de recherche évaluative, une démarche analytique susceptible de mobiliser différentes méthodes ou outils d'évaluation. Le terme « théorie » est ici à comprendre comme la décomposition et l'explicitation d'une chaîne causale liant l'intervention publique et les résultats et impacts escomptés sur les publics ciblés et bénéficiaires. Celle-ci prend alors la forme d'une

représentation graphique qui devient alors « l'échafaudage » sous-tendant les investigations de l'évaluation. Selon les usages on peut nommer cet outil « logique d'intervention », « diagramme logique d'impact (DLI) », « logigramme », « théorie du changement » ou « chemin d'impact ». Cette démarche est largement répandue dans la pratique évaluative, cependant elle recouvre une diversité de pratiques de qualité parfois variable et mobilisant des approches plus ou moins techniques.

Une approche basée sur la théorie, contrairement à ce que pourrait évoquer cette appellation, va chercher à être au plus près de la réalité du terrain et du contexte de l'intervention. Le terme théorie prend ici un sens plus ou moins scientifique selon les usages. A minima, une approche basée sur la théorie s'intéresse à la théorie de mise en œuvre d'une intervention, c'est-à-dire aux actions que l'on met en place pour la bonne réalisation de l'intervention permettant d'atteindre les premiers résultats (logique d'intervention). Dans la plupart des cas, elle tend à faire un pas de plus et à expliciter les liens causaux hypothétiques entre des mécanismes délivrés par une intervention et leurs résultats escomptés allant jusqu'aux impacts finaux (théorie du changement). Cette analyse est généralement structurée en plusieurs catégories :

- Les réalisations qui décrivent quelles actions sont mises en par les autorités publiques;
- Les résultats qui décrivent les premiers effets immédiats directement liés aux réalisations et pouvant être observés sur les publics participants aux actions;
- Les impacts intermédiaires et finaux qui décrivent les impacts que devraient avoir les réalisations sur les bénéficiaires finaux dont on cherche à améliorer la situation.

La conduite de cette démarche débute systématiquement par un travail d'élaboration de la « théorie ». Celle-ci s'appuie généralement sur une analyse documentaire ainsi que des discussions avec les parties

prenantes, afin d'identifier les composantes clés de l'intervention, les publics ciblés, les principaux résultats escomptés ainsi que les impacts finaux ciblés par l'intervention. La théorie est ainsi souvent co-construite entre les équipes d'évaluation et les parties prenantes et présente graphiquement les principales étapes, allant des réalisations aux résultats et impacts finaux en les connectant entre eux et en explicitant les hypothèses pour passer d'une étape à une autre. Ces hypothèses peuvent couvrir les conditions de succès, les risques liés au contexte, mais aussi dans le cadre d'une évaluation réaliste les mécanismes psycho-sociaux à l'œuvre ou dans le cadre d'une analyse de contribution, préciser les hypothèses de contribution de l'intervention. À ce stade, le travail peut être consolidé par une analyse de la littérature et donc une mobilisation de théories cette fois académiques pour étayer certains liens logiques (par exemple dans une intervention ciblant le retour à l'emploi, informer une relation causale entre un niveau de formation et emploi).

Ensuite, cette théorie devient l'échafaudage de l'évaluation. Deux types d'investigations sont alors conduites :

- D'abord une analyse du changement : est-ce que les étapes identifiées ou souhaitées lors de la conception de la théorie se vérifient sur le terrain? Pour qui et dans quels contextes?
- Puis une analyse causale : Comment peut-on expliquer le passage ou non d'une étape à une autre? Dans quelle mesure l'intervention contribue-t-elle à générer ces résultats? Dans quelles conditions? Quels mécanismes psychosociologiques peuvent l'expliquer?

Les questions d'évaluation sont donc structurées autour de vérification de la théorie et de son explication. Dans cette tâche, l'équipe d'évaluation peut mobiliser une diversité de méthodes ou d'outils d'évaluation. Par exemple, on peut avoir recours à des méthodes d'estimation quantitatives des résultats ou des impacts et les articuler avec des méthodes qualitatives pour approfondir l'analyse causale.

In fine, les résultats de cette démarche prennent la forme d'une nouvelle théorie « vérifiée » de l'intervention montrant les processus causaux effectivement à l'œuvre, ceux ne fonctionnant pas comme prévu, et apportant des explications sur ces phénomènes. Ce type de résultat permet alors d'identifier à quelle étape, dans quel contexte et auprès de quels publics l'intervention a rencontré des difficultés et de proposer des pistes d'améliorations opérationnelles et stratégiques pour les prochaines programmations.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

Mobiliser une théorie du changement dans une évaluation peut servir différents usages. Décomposer les étapes et les liens de cause à effet permet lors des investigations d'identifier quelles sont les étapes fonctionnelles de la théorie et celles plus problématiques. Une intervention publique n'est jamais complètement un échec ou un succès. Par exemple, si l'on constate que l'intervention n'a pas atteint les résultats souhaités, si l'on renseigne les différentes étapes de la théorie du changement lors de l'évaluation, on peut identifier si les difficultés sont liées à des problèmes liés à la mise en œuvre (des actions qui n'ont pas été correctement organisées ou une difficulté à mobiliser les publics ciblés) ou la capacité de l'intervention à avoir les effets escomptés (les actions ont bien été organisées et ont atteint les publics ciblés, mais ne parviennent pas à générer les effets escomptés sur tous les publics ou certaines catégories). Il est alors possible d'approfondir l'analyse de la causalité pour tenter de comprendre si ces difficultés sont liées au programme lui-même ou à des éléments de contextes extérieurs qui n'ont pas suffisamment été pris en compte lors de la conception ou la mise en œuvre du programme (concurrence avec d'autres dispositifs, contraintes freinant la participation des publics, mauvaise analyse des besoins, ou changement d'une conjoncture économique par exemple).

Cette démarche permet de répondre aux questions de type : « Dans quels cas l'intervention X a-t-elle contribué à générer les résultats Y et pourquoi? ». Plus globalement elle peut être utilisée pour répondre à l'ensemble des registres de questionnements évaluatifs sur l'efficacité, l'impact, la pertinence, la cohérence interne ou externe ainsi que l'efficience. En effet, chacun de ces registres peut être investigué pour expliquer le succès ou l'échec dans la progression d'une étape à une autre de la théorie.

Elle peut être utilisée lors de la conception d'une intervention pour notamment réunir les parties prenantes, co-construire cette théorie et en anticiper les risques en amont de la mise en œuvre et modifier la programmation pour renforcer les chances de succès de l'intervention. Dans ce cas, elle est d'ailleurs utile à l'appropriation d'une vision commune et partagée de l'intervention par les parties prenantes.

Elle est aussi particulièrement pertinente dans les évaluations chemin faisant ou *in itinere* afin de vérifier au cours de la mise en œuvre l'atteinte des différentes étapes et d'ajuster la mise en œuvre du projet pour favoriser l'atteinte des résultats. Enfin, dans le cas d'une évaluation ex post, en fin de projet, elle permet de comparer la théorie initiale planifiée avec la théorie vérifiée dans le cadre de l'évaluation et de comprendre les raisons de ces écarts.

III. Un exemple de l'utilisation de cette approche : évaluation d'un programme de nutrition

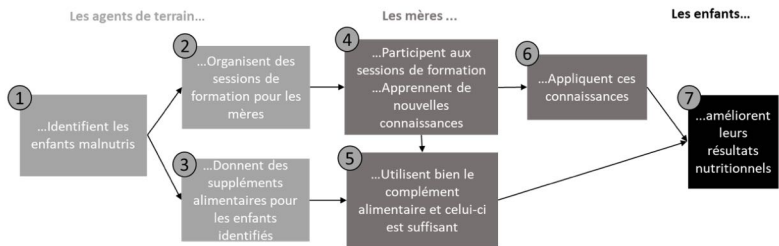
Les approches d'évaluation basées sur la théorie sont mobilisées dans une diversité de contextes et sont autant adaptées aux projets simples qu'aux politiques publiques complexes. Un exemple de son utilisation dans le cadre d'un projet simple peut illustrer brièvement la valeur ajoutée de cette approche.

Dans un document pour l'International Initiative for Impact Evaluation, Howard White (2009) présente le cas de l'évaluation du projet de nutrition intégrée au Bangladesh (BINP). Ce projet repère les enfants en retard de croissance et propose une distribution de suppléments alimentaires ainsi que des conseils de nutrition aux mères de ces enfants. L'objectif final est une augmentation de la croissance des enfants.

Une première évaluation par groupe de comparaison par appariement de scores de propension n'a pas constaté d'impact de ce projet sur l'état nutritionnel des enfants, mais un impact positif uniquement sur les enfants les plus malnutris. Néanmoins, ce résultat en tant que tel ne permet pas d'expliquer ce résultat, de disposer d'enseignements sur les dysfonctionnements du projet ni sur les mesures pouvant être prises pour l'améliorer.

Une évaluation complémentaire basée sur la théorie a permis d'enrichir ces résultats et de proposer des orientations pour l'action. Cette démarche s'est d'abord intéressée à reconstituer la théorie du projet puis à expliciter et investiguer les hypothèses causales sous-tendant cette action.

Figure 1 : Théorie du projet de nutrition intégrée (adaptation par l'autrice de cette fiche)



Une première étape est d'abord la bonne identification des enfants malnutris (1). Cette étape s'appuie sur l'hypothèse que les parents amènent bien leurs enfants aux centres de prévention et ensuite que les enfants malnutris soient bien repérés. Ici il semble que le programme parvenait à atteindre le public ciblé (90% des femmes admissibles ont amené leurs enfants), néanmoins la sélection des enfants par les agents communautaires a montré plusieurs erreurs de type 1 (des enfants malnutris non sélectionnés) et de type 2 (des enfants sélectionnés alors qu'ils ne sont pas malnutris).

La chaîne causale 2-4-6-7 s'appuie sur l'hypothèse que les bons publics cibles ont été identifiés et que les modes d'action sont pertinents. Or, une étude anthropologique et des focus groupes ont ici révélé que les mères ciblées par ce dispositif n'avaient finalement que peu d'influence sur la nutrition de leur enfant, car les hommes avaient la responsabilité des courses alimentaires et pour la plupart les belles-mères étaient chargées de la préparation des repas. Ainsi, bien qu'elles participent aux formations, et apprennent de nouvelles connaissances les mères n'étaient en position de les appliquer que lorsqu'elles seraient belles-mères à leur tour.

La chaîne causale 3-5-7 s'appuie sur l'hypothèse que les enfants malnutris ont effectivement été sélectionnés, et que le complément alimentaire est bien utilisé en complément et non en substitution ou partagé entre les enfants et que celui-ci est bien suffisant pour améliorer les résultats nutritionnels des effets. Or il s'avère que les enquêtes menées ont montré que ces deux hypothèses n'étaient pas vérifiées dans de nombreux cas.

Cette évaluation a permis d'identifier des recommandations très opérationnelles pour améliorer les effets du projet notamment : La participation des belles-mères et des maris aux séances de conseils en nutrition ainsi qu'une meilleure formation des agents communautaires en charge de la sélection des enfants et un ciblage sur les enfants les plus malnutris.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

L'analyse de la qualité de la mobilisation de cette approche peut s'effectuer à deux moments de l'évaluation.

D'abord, la phase d'élaboration de la théorie doit aboutir à la théorie la plus plausible avant l'investigation de terrain. Dans l'idéal une bonne théorie est co-construite entre l'équipe d'évaluation et les différentes parties prenantes du dispositif et mobilise aussi des connaissances extérieures pouvant éclairer la plausibilité des hypothèses de liens de causalité entre les différentes étapes de cette théorie. Être plausible signifie qu'elle rend compte d'une ambition raisonnable de l'intervention dans la mesure des changements qu'elle est susceptible de générer. Enfin, une bonne théorie reste synthétique, simple à comprendre et indique clairement par une représentation graphique qui agit et qui est susceptible de changer de comportement ou de voir sa situation s'améliorer suite à la contribution de l'intervention étudiée.

Ensuite, les critères de qualité de la phase de test de la théorie s'inscrivent dans les standards généraux de la collecte et de l'analyse de données. Les outils de collecte doivent être diversifiés pour rendre compte d'une pluralité de sources de données (quantitatives et qualitatives) et de points de vue afin de consolider au maximum les résultats. La collecte et l'analyse sont idéalement conduites de manière itérative (par phases successives) afin d'affiner la théorie du changement et ses hypothèses tout au long de l'évaluation.

Les résultats de l'évaluation doivent ensuite être présentés autour de la vérification de la théorie et présenter ce qui a marché ou non pour quels types de publics pourquoi et dans quelles conditions.

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

Le principal atout de cette approche est « d'ouvrir la boîte noire » de l'action publique. À la différence des approches expérimentales ou quasi-expérimentales qui permettent d'estimer des impacts sans analyser leur processus de production, l'évaluation basée sur la théorie permet d'identifier et de décomposer les mécanismes conduisant à la production de l'impact. Elle applique une démarche analytique qui décompose l'intervention en plusieurs étapes et en liens de causalité et qui permet de distinguer ce qui relève du contexte ou de l'intervention. Expliciter les hypothèses sous-jacentes à l'action oblige l'équipe d'évaluation à s'intéresser aux contextes de mise en œuvre, aux caractéristiques des publics ainsi qu'aux « paris » de l'intervention c'est-à-dire la manière dont elle est susceptible d'être un levier de changement. Les résultats d'évaluation sont donc particulièrement contextualisés et génèrent des enseignements même lorsque l'intervention n'est pas finalisée puisque chaque étape de la mise en œuvre peut être investiguée.

Cependant, cette approche peut être déstabilisante, car contrairement à d'autres méthodes, il n'y a pas une manière et un protocole distinct à suivre pour la mettre en œuvre. Ainsi, plusieurs praticiens ou praticiennes de l'évaluation peuvent revendiquer s'inscrire dans cette démarche sans pour autant correspondre aux standards de qualité de cet exercice.

Par ailleurs, on a pu reprocher à cette méthode d'accorder trop d'attention à l'analyse de la théorie escomptée et d'occulter l'identification d'explications alternatives des processus observés ou de résultats non anticipés. C'est en effet un écueil auquel l'équipe d'évaluation doit être attentive en tentant à la fois lors de la conception de la théorie et de sa vérification d'être aussi ouvert à la formulation et l'identification de ces explications alternatives. La mobilisation de la méthode d'analyse de

contribution permet notamment de concentrer l'analyse sur la place de l'intervention dans un paquet causal plus complexe et d'investiguer les explications alternatives.

Quelques références bibliographiques pour aller plus loin

Birckmayer, Johanna D. et Weiss, Carol H.. 2000. « Theory-Based Evaluation in Practice, What Do We Learn? » *Evaluation Review* 24 (4): 407-431.

Funnell, Sue. et Rogers, Patricia. 2011. *Purposeful Program Theory, Effective use of Theories of Change and Logic Models*. Jossey-Bass, Chichester.

Rogers, Patricia. et Petrosino, Anthony. et al.. 2000. « Program Theory Evaluation: Practice, Promise, and Problems ». *New directions for evaluation* 87, 5-13.

Van Es, Marjan. et Guijt, Irene. et Vogel, Isabel. 2015. *Theory of change thinking in practice: A stepwise approach*, Hivos ToC Guidelines, Hivos people unlimited.

Weiss, Carol. 1997. « Theory-based Evaluation: Past, Present, and Future ». *New directions for evaluation* 76, 41-55.

White, Howard. 2009. « L'évaluation d'impact basée sur la théorie : principes et pratiques », *Working paper* 3, International initiative for Impact Evaluation.

2I. Évaluation réaliste

SARAH LOUART, HABIBATA BALDÉ, ÉMILIE ROBERT ET VALÉRY RIDDE

Résumé

L'évaluation réaliste est fondée sur une conception des politiques publiques comme des interventions qui produisent leurs effets par le biais de mécanismes qui ne se déclenchent que dans des contextes spécifiques. L'analyse de ces liens entre des contextes, des mécanismes et des effets est donc au cœur de cette approche. Celle-ci peut prendre appui sur une diversité de méthodes, mais va dans tous les cas avoir recours à des méthodes qualitatives pour investiguer les mécanismes en jeu. Relevant de la famille des évaluations basées sur la théorie, l'évaluation réaliste aspire à produire des théories de moyenne portée permettant notamment de faciliter le transfert des connaissances produites sur l'intervention étudiée à d'autres contextes ou d'autres interventions du même type.

Mots-clés : Méthodes qualitatives, évaluation basée sur la théorie, configurations contexte – mécanisme – effet, théorie de moyenne portée, réalisme critique

I. En quoi consiste cette approche?

Le réalisme critique, courant de pensée porté notamment par Roy Bhaskar (1975), a ouvert la voie au développement de l'approche réaliste pour l'évaluation. Dans son ouvrage *A Realist Theory of Science*, il se demande : comment doit être le monde pour que la science soit possible? L'idée est de se questionner sur la nature de la réalité, puisque c'est ce qui va ensuite déterminer la manière dont on peut l'explorer et l'appréhender.

Bhaskar avance qu'il existe des structures, des pouvoirs, des mécanismes (par exemple la gravité) qui existent et qui peuvent produire des effets, même si nous ne les connaissons pas. Une feuille peut tomber de l'arbre et atteindre le sol sous l'effet de la gravité, même si nous ne l'avons pas observée. L'objectif est alors d'essayer de produire des connaissances sur les mécanismes, pouvoirs, structures qui existent et sur leurs façons d'agir, notamment les conditions qui favorisent leur déclenchement et les effets qu'ils peuvent produire. Ces théories sont donc des « énoncés sur la façon dont les choses agissent dans le monde » (Bhaskar 1975). L'objectif des chercheur·e·s est alors de produire des théories qui vont essayer d'élucider l'existence des mécanismes et leurs modes de fonctionnement. Cependant, ces théories seront toujours perfectibles et évolutives, la réalité et les connaissances évoluent. Pour les chercheur·e·s « réalistes », il y a une réalité, mais il n'y a pas de vérités générales valables en tout temps et en tout lieu.

L'approche réaliste de l'évaluation des politiques publiques se base sur ces principes (Pawson et Tilley 1997; Westhorp et al. 2011; Westhorp 2014; Robert et Ridde 2020). Elle a été introduite par Pawson et Tilley (1997). Le présupposé est que les politiques ont de réels effets, mais elles ne les provoquent pas directement. Elles permettent, ou non, le déclenchement de mécanismes, qui vont, eux-mêmes, produire des effets. Les mécanismes sont la réaction des personnes aux ressources, aux sanctions ou aux opportunités (selon le type de politique publique) mises à leur disposition dans le cadre de la politique (Lacouture et al. 2015). Les acteurs sont donc les moteurs du changement, ce sont leurs réactions qui vont produire des effets, qu'ils soient positifs, négatifs, attendus ou inattendus. Il ne suffit donc plus de répondre à la question : la politique fonctionne-t-elle (ou non), mais plutôt, d'investiguer les mécanismes qui, déclenchés par la politique dans des contextes spécifiques, vont produire des effets. Cela va permettre de répondre à la question : comment la politique fonctionne-t-elle ou non, pourquoi, pour qui, et dans quels

contextes? L'objectif de l'évaluation réaliste est donc de faire des liens entre le déclenchement de mécanismes et des facteurs contextuels, et entre l'action de ces mécanismes et la survenue d'effets.

Pour répondre à toutes ces questions évaluatives, l'évaluation réaliste va mobiliser une approche d'évaluation basée sur la théorie (cf chapitre séparé). Il s'agit de partir de la théorie de l'intervention de la politique à évaluer (la façon dont on s'attend à ce que la politique produise ses effets), et de la faire évoluer, en fonction des connaissances existantes et des données de terrain. L'objectif est d'arriver à formuler ce qu'on appelle une théorie de moyenne portée (Merton 1968). Il s'agit d'une théorie qui se situe entre la théorie de l'intervention de la politique et une théorie qui se voudrait générale. C'est une théorie explicative d'une régularité (tendance observée) contextualisée (liée à des contextes particuliers). Pour arriver à cela, différentes étapes de l'évaluation sont à organiser.

(1) Reconstruire la théorie d'intervention

Il s'agit de commencer par comprendre la politique mise en œuvre. Il y a tout un ensemble de croyances ou de suppositions qui soutiennent les activités de la politique. En effet, toute politique repose sur une théorie du changement, c'est-à-dire une idée selon laquelle les activités mises en œuvre peuvent produire des changements. Il s'agit de répondre à des questions telles que : quelles sont les ressources mises à disposition par la politique et pourquoi? Quels changements la politique pourrait-elle générer et comment? Quels éléments du contexte pourraient avoir une influence sur la politique? Cette théorie de l'intervention n'est initialement souvent pas explicite, et il s'agit de l'investiguer. Cela peut passer par des discussions avec l'équipe de mise en œuvre et par une revue des documents. Souvent utilisés de façon interchangeable, les concepts de « théorie de l'intervention » et de « théorie du changement » présentent toutefois des différences. Ainsi la théorie de l'intervention est

une théorie détaillée, centrée sur l'intervention, et décrivant l'ensemble des composantes d'une intervention et leur relation logique jusqu'à leur impact souhaité. Dans une perspective réaliste, elle intègre les concepts de mécanisme et de contexte. La théorie du changement est, quant à elle, centrée sur l'objectif de l'intervention et le changement social qu'elle vise : elle décrit ainsi la partie du raisonnement logique entre les résultats attendus et les impacts souhaités d'une intervention.

(2) Formuler des propositions théoriques sur la façon dont la politique peut produire des effets

À cette étape, il faut se baser à la fois sur la théorie de la politique reconstituée à l'étape précédente, et sur les connaissances scientifiques. L'objectif est de formuler des propositions théoriques sur la façon dont les activités de la politique, dans des contextes spécifiques, pourraient déclencher des mécanismes, qui pourront eux-mêmes produire des effets, et lesquels. Il s'agit de travailler sur les interactions entre un contexte particulier, la possibilité de déclenchement d'un mécanisme via la politique, et la production d'un effet. Le contexte est l'ensemble des facteurs extérieurs à la politique qui ont une influence sur le déclenchement d'un mécanisme (ex : caractéristiques socio-économiques des participant-e-s, normes sociales, relations interpersonnelles, environnement politique, etc.). Les effets sont les résultats produits par le déclenchement de ces mécanismes. Ces interactions sont donc des propositions théoriques sur la façon dont la politique devrait fonctionner et pourquoi.

(3) Tester empiriquement les propositions théoriques

Sur la base des propositions théoriques construites *a priori*, l'objectif est ensuite de les mettre à l'épreuve des faits, afin de les confirmer, réviser ou préciser. Cela permet de faire évoluer les propositions théoriques et donc la théorie de la politique. Cette mise à l'épreuve des faits pourra mobiliser des méthodes de collecte des données à la fois quantitatives et qualitatives. En effet, l'évaluation réaliste ne repose pas forcément sur un seul type de méthode ou de données. L'objectif est d'utiliser une panoplie de méthodes en fonction de leur pertinence pour tester les propositions théoriques. Il ne s'agit donc pas vraiment d'une méthode de recherche mais plutôt d'une « logique d'enquête » (Pawson et Tilley 2004). Néanmoins, pour investiguer les mécanismes et donc les raisonnements des acteurs et actrices impliqué-e-s dans la politique, il est nécessaire de mobiliser au moins des données de nature qualitative pour comprendre comment les acteurs et actrices perçoivent et réagissent à la politique. L'objectif de cette étape est de repérer, dans les données issues du terrain, des liens entre des contextes, des mécanismes et des effets, qui surviennent de façon régulière.

(4) Préciser la théorie de moyenne portée

À partir des données de terrain, les propositions théoriques formulées sont donc testées et précisées. Les propositions initialement construites évoluent et sont complétées avec les nouvelles données. Cela nous permet d'avoir une théorie consolidée, qui peut être formalisée par un ensemble de configurations « contexte, mécanismes, effets » (CME), pour expliquer comment, pourquoi, pour qui et dans quels contextes ce type de politique peut fonctionner ou non. Comme les mécanismes sont déclenchés par la politique, et sont forcément liés à une personne ou groupe de personnes, il est possible de formuler les théories sous forme

de configurations plus complètes : ICAME (intervention – contexte – acteur – mécanisme – effet). Cette théorie est appelée théorie de moyenne portée car elle a une validité plus large que la théorie de la politique, qui est très spécifique. La théorie de moyenne portée est plus abstraite et peut servir de base pour analyser et évaluer d'autres politiques du même type.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

L'évaluation des politiques publiques a pour objectif d'orienter l'action publique en identifiant les activités les plus pertinentes à engager au vu des objectifs souhaités. Pour cela, il est nécessaire de s'appuyer sur les leçons apprises de la mise en œuvre de politiques passées ou en cours. Ces leçons ne doivent pas se résumer à évaluer la réalisation ou non des objectifs d'une action. Elles doivent également se nourrir de la mise à l'épreuve des hypothèses et préjugés qui ont fondé cette politique, ainsi que des stratégies d'action de celles et ceux chargé-e-s de la mettre en œuvre. L'approche réaliste permet donc de se distinguer des approches plus traditionnelles de l'évaluation, qui visent souvent à évaluer uniquement l'efficacité d'une politique, à l'aide d'indicateurs plutôt quantitatifs. Souvent, ces méthodes, à elles seules, se révèlent insuffisantes pour tirer des leçons pertinentes de la mise en œuvre de politiques complexes. Or, la plupart des politiques sont complexes car elles interviennent à différents niveaux d'action, impliquent une multitude de personnes, se transforment au fur et à mesure de leur mise en œuvre, sont influencées par le contexte et une myriade de facteurs. Il est donc utile de se tourner vers d'autres approches, comme l'évaluation réaliste, qui permettent de prendre en compte la complexité des politiques. Elle permet de saisir comment une politique peut, ou non, engendrer des changements. C'est une recherche évaluative de type explicative.

Il est souvent mis en avant dans la littérature critique sur les politiques publiques, qu'un même type de politique est diffusé sans être adapté dans d'autres contextes (Olivier de Sardan 2021). Pourtant, cette diffusion de politiques standardisées ne permet souvent pas de produire les mêmes résultats ailleurs. L'approche réaliste permet d'expliquer cela et peut permettre d'éviter de tels écueils. Elle aide à comprendre ce qui favorise, ou non, le déclenchement des mécanismes qui produisent des effets positifs, et à comprendre comment et pourquoi ces déclenchements pourraient potentiellement survenir ailleurs. Comprendre pourquoi et comment les politiques publiques fonctionnent, auprès de quels bénéficiaires et dans quels contextes, permet de fournir des orientations à la prise de décision. Se poser la question de « pour qui » la politique fonctionne est également une interrogation essentielle dans le cadre de l'évaluation des politiques publiques. Cela est nécessaire pour tenir compte des retombées différentes de la politique pour les différents sous-groupes, notamment les plus marginalisés, chez lesquels on peut observer des effets différentiels, contre-intuitifs voire indésirables. Tous ces questionnements, que l'on retrouve dans l'approche réaliste, peuvent donner des indications sur la pertinence de mettre en œuvre une politique dans un autre contexte, ou sur la façon dont on peut adapter ou faire évoluer la politique pour qu'elle ait le plus de chances de produire les effets escomptés. Elle est donc une approche particulièrement adaptée quand la politique vise à être mise à l'échelle et étendue à d'autres populations, dans d'autres contextes.

Un raisonnement réaliste, s'appuyant sur les résultats d'évaluations antérieures ou sur une revue des écrits scientifiques selon cette approche, peut tout à fait être utilisé pour guider la formulation d'une politique avant sa mise en œuvre. Néanmoins, l'évaluation réaliste de la politique ne pourra pas être réalisée uniquement *a priori* si aucune donnée sur des résultats n'est disponible, puisque pour élaborer les configurations CME, des données sur les effets sont nécessaires.

III. Un exemple d'utilisation de cette approche pour évaluer la mise en œuvre de la couverture sanitaire universelle

La couverture sanitaire universelle (CSU) promeut l'accès de toutes les personnes aux services de santé dont elles ont besoin, sans être exposées à des difficultés financières. Pour atteindre ce but, l'Organisation mondiale de la santé (OMS) a mis en place un Partenariat pour soutenir la CSU dans plusieurs pays. Ce partenariat vise à soutenir un dialogue collaboratif sur les politiques, instruments de gouvernance dans les pays qui ont pour objectif de mettre en œuvre des actions pour la CSU. Cette intervention consiste donc à mettre à disposition des ressources et expertises (ex : assistant·e technique, formation pour les cadres des ministères, etc.) en fonction des besoins des Ministères de la santé (Robert et Ridde 2020). Ce type d'intervention est donc complexe, prend place dans des contextes très différents et sous des formes variables. Elle ne peut pas être évaluée en mobilisant uniquement des données et indicateurs de types quantitatifs. Par contre, l'évaluation réaliste a permis de mieux comprendre comment ce Partenariat peut fonctionner, et les potentielles différences de résultats selon les contextes de mises en œuvre. L'objectif général de l'étude était donc de comprendre comment, dans quels contextes et via quels mécanismes le Partenariat peut soutenir le dialogue sur les politiques.

L'objectif était d'investiguer : 1) comment et dans quels contextes le Partenariat peut-il initier et nourrir le dialogue sur les politiques? 2) comment la dynamique de collaboration se déroule-t-elle au sein du dialogue sur les politiques soutenu par le Partenariat? (Robert et al. 2022) Une étude de cas multiples a été menée dans six pays. Des propositions théoriques sur la façon dont la politique pouvait fonctionner ont pu être tirées des documents du projet mais aussi des théories existantes dans les écrits scientifiques, par exemple les théories sur les relations partenariales et la gouvernance collaborative. Un exemple de proposition

théorique formulée est que le renforcement des capacités (par la formation, l'expertise technique et le soutien continu de l'OMS) donnerait du pouvoir au ministère de la santé (M) tout en déclenchant une compréhension commune de la gouvernance et du dialogue sur les politiques (M) ; ce qui devrait amener le ministère de la Santé à mener des dialogues sur les politiques inclusifs et participatifs (E). Le déclenchement de ces mécanismes pourrait être favorisé par des facteurs contextuels, comme par exemple le fait que l'OMS et le ministère de la Santé entretiennent des relations durables (C) ou que les ressources humaines des deux institutions qui participent au Partenariat sont stables (C).

Une démarche collaborative a été adoptée, en impliquant les parties prenantes aux étapes clés de l'évaluation : développement du protocole, élaboration de la théorie d'intervention, interprétation des résultats, etc. En s'appuyant à la fois sur les théories afin de monter en abstraction, la théorie d'intervention et les données de terrain pour consolider ou réfuter les propositions théoriques initiales, plusieurs configurations CME ont pu être formulées. Par exemple : le partenariat facilite le lancement du dialogue sur les politiques (E) en suscitant l'intérêt des parties prenantes pour une collaboration multisectorielle (M), à condition que ces dernières reconnaissent leur interdépendance et l'incertitude qui pèse sur la gestion des problèmes de santé essentiels (C). On voit qu'un des effets attendus par l'OMS lors de la mise en place du Partenariat ne se réalisera que dans un contexte particulier qui permettra le déclenchement d'un mécanisme réactionnel spécifique de la part des parties prenantes. Ce type de résultats pourra soutenir la mise en œuvre des actions similaires, aider à l'adapter ou à connaître les contextes où ce type d'intervention a le plus de chances de répondre positivement aux objectifs attendus.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

L'évaluation réaliste est plus une approche de l'évaluation qu'une méthode. Elle fait partie de la catégorie de « recherche évaluative ». Pour juger de la qualité d'une évaluation réaliste, il faut donc plutôt s'assurer que l'évaluation respecte certains critères de base de l'approche. Par exemple, l'évaluation doit se focaliser sur la découverte des mécanismes à l'œuvre, et le concept de mécanisme doit être correctement compris et appliqué. Il est nécessaire de retrouver dans le processus d'évaluation la mise en relation de contextes, mécanismes, effets, et acteurs, par la présence de configurations. L'évaluation doit permettre de générer une plus grande abstraction par rapport à la théorie de la politique. D'autres éléments peuvent également favoriser une évaluation réaliste de qualité : la réalisation d'une revue des écrits scientifiques pour investiguer les théories existantes et soutenir la formulation des propositions théoriques à tester; le croisement des sources de données (qualitatives et quantitatives); l'implication des différentes parties prenantes de la politique aux différentes étapes de l'évaluation, etc. Il existe des guides, tels que les « *Quality Standards for Realist Evaluation* » (Wong et al. 2016) qui fournissent une orientation à chaque étape de l'évaluation, afin de réaliser une évaluation réaliste de qualité.

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

L'évaluation réaliste a de nombreux atouts. Elle permet de prendre en compte la complexité des politiques publiques, ainsi que celle du monde social, politique, économique, dans lequel elles prennent place. Elle s'appuie sur une démarche collaborative, et favorise l'implication de toutes les parties prenantes (au niveau institutionnel, opérationnel et

réceptionnaire de la politique). Plus particulièrement, cette démarche permet de mettre les « bénéficiaires » et personnes de premières lignes au centre de l'évaluation, en les considérant comme des expert·e·s. Ce sont leurs réactions que l'on essaie de comprendre, et ce sont elles et eux qui sont les plus à même de nous informer.

Elle permet d'expliquer des processus et résultats multiples, de mettre en évidence les résultats inattendus des politiques, et de répondre à des questions évaluatives souvent négligées (comprendre le comment plutôt qu'uniquement le résultat). Le fait de chercher à comprendre de façon approfondie le fonctionnement des politiques permet de fournir des connaissances qui sont plus à même d'être mobilisables dans d'autres contextes. Son attention au contexte est fondamentale car trop souvent oubliée dans les approches évaluatives standards. De plus, son ancrage dans la théorie permet de favoriser l'utilisation et le cumul des connaissances disponibles. Elle permet de mobiliser concrètement les connaissances scientifiques, alors qu'elles sont souvent trop peu utilisées sur le terrain. Le fait de s'appuyer à la fois sur les écrits scientifiques et sur les données issues du terrain, permet de s'assurer d'une certaine transférabilité des résultats produits et de fournir des recommandations appropriées aux personnes décisionnaires politiques (pertinence ou non de mettre en œuvre ce type de politique dans certains contextes, comment adapter une politique à un contexte spécifique, etc.). Enfin, elle permet la collaboration entre des équipes de travail ayant des expertises et des domaines de recherche différents, ainsi que la mobilisation de méthodes de recherche très différentes.

Néanmoins, mobiliser cette approche comporte quelques défis. D'abord, elle demande du temps et elle peut être difficile à maîtriser. Les concepts de contexte et de mécanisme peuvent être difficiles à appréhender et à opérationnaliser. Il y a encore un manque d'enseignements dédiés à la pratique de l'évaluation avancée et de la recherche évaluative dans lesquels on enseigne l'évaluation réaliste. De plus, beaucoup de parties prenantes de l'évaluation (bailleurs, partenaires opérationnels, comités d'éthique, etc.) ne connaissent pas cette approche, ce qui peut poser

des problèmes de compréhension sur ce qu'il est possible ou non de faire, et donc de répondre à leurs attentes. Ensuite, l'évaluation reste très marquée par la recherche de résultats d'impact mesurés à l'aide d'indicateurs. Ici, c'est un tout autre format de résultats qui pourra être fourni aux commanditaires et aux parties prenantes intéressées. Enfin, cette approche de l'évaluation n'est pas simple, et ne permet pas de produire des résultats linéaires : plusieurs mécanismes peuvent agir en même temps, avoir des influences contraires sur les résultats ; un effet dans une configuration peut devenir un contexte dans une autre. Les configurations CME peuvent donc parfois être difficiles à construire.

Quelques références bibliographiques pour aller plus loin

Bhaskar, Roy. 1975. *A Realist Theory of Science*.

Lacouture, Anthony. et Breton, Eric. et Guichard, Anne. et Ridde, Valéry. 2015. « The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation. » *Implementation Science* 10 (1): 153. <https://doi.org/10.1186/s13012-015-0345-7>.

Merton, Robert C.. 1968. *Social Theory and Social Structure*. Simon and Schuster.

Olivier de Sardan, Jean-Pierre. 2021. *La Revanche des contextes: Des mésaventures de l'ingénierie sociale en Afrique et au-delà*. Paris: Karthala.

Pawson, Ray. et Tilley, Nicholas. 1997. *Realistic Evaluation*. London: Sage.

Pawson, Ray. et Tilley, Nicholas. 2004. *Realist Evaluation*. DPRN Thematic Meeting 2006 Report on Evaluation.

- Robert, Emilie. et Ridde, Valéry. 2020. *Dealing With Complexity and Heterogeneity in a Collaborative Realist Multiple Case Study in Low- and Middle-Income Countries*. SAGE Research Methods Cases. SAGE Publications Ltd. <https://doi.org/10.4135/9781529732306>.
- Robert, Emilie. et Ridde, Valéry. et Rajan, Dheepa. et Sam, Omar. et Dravé, Mamadou. et Porignon, Denis. 2019. « Realist Evaluation of the Role of the Universal Health Coverage Partnership in Strengthening Policy Dialogue for Health Planning and Financing: A Protocol. » *BMJ Open* 9 (1): e022345. <https://doi.org/10.1136/bmjopen-2018-022345>.
- Robert, Emilie. et Zongo, Sylvie. et Rajan, Dheepa. et Ridde, Valéry. 2022. « Contributing to Collaborative Health Governance in Africa: A Realist Evaluation of the Universal Health Coverage Partnership. » *BMC Health Services Research* 22 (1): 753. <https://doi.org/10.1186/s12913-022-08120-0>.
- Westhorp, Gill. 2014. « Realist Impact Evaluation: An Introduction ». Methods Lab.
- Westhorp, Gill. et Prins, Ester. et Kusters, Cecile. et Hultink, Mirte. et Guijt, Irene. et Brouwers, Jan. 2011. « Realist Evaluation: An Overview ». Seminar report.
- Wong, Geoff. et Westhorp, Gill. et Manzano, Ana. et Greenhalgh, Joanne. et Jagosh, Justic. et Greenhalgh, Trish. 2016. « RAMESES II reporting standards for realist evaluations ». *BMC Medicine* 14 (1): 96. <https://doi.org/10.1186/s12916-016-0643-1>.

22. L'analyse de contribution

THOMAS DELAHAIS

Résumé

L'analyse de contribution est une approche évaluative « basée sur la théorie » particulièrement adaptée à l'évaluation d'interventions complexes. Elle consiste à formuler progressivement des « hypothèses de contribution », dans un processus impliquant les parties prenantes de la politique, pour ensuite tester ces hypothèses de façon systématique à partir d'une diversité de méthodes (qui peuvent être qualitatives ou mixtes).

Mots-clés : Méthodes mixtes, interventions complexes, hypothèses de contribution, démarche abductive, contexte, chemins d'impact, paquets causaux, approche narrative

I. En quoi consiste cette approche?

L'analyse de contribution est une approche évaluative dite « basée sur la théorie » (EBT) : elle s'organise autour d'un processus visant à 1) élaborer un ensemble d'hypothèses portant sur les effets d'une intervention évaluée (comment ces effets sont obtenus, dans quels cas, pourquoi...) – appelé « théorie du changement »; puis à 2) tester ces hypothèses à travers la collecte et l'analyse d'informations empiriques; pour enfin 3) mettre à jour la théorie initiale en indiquant quelles hypothèses sont vérifiées.

Tout comme l'évaluation réaliste ou le traçage de processus (*process tracing*, voir chapitres dédiés dans cet ouvrage), par exemple, l'analyse de contribution fait partie des EBT de nouvelle génération qui ont émergé dans les années 2000 (elles sont parfois réunies sous l'appellation d'évaluations d'impact basées sur la théorie – EIBT). Elle aborde les interventions évaluées comme des objets complexes dans des environnements complexes. Ainsi, l'analyse de contribution récuse fondamentalement l'idée que les interventions puissent intrinsèquement « marcher »; le succès ou l'échec dépend toujours d'une pluralité de facteurs et de contextes, que l'évaluation doit documenter. En cela, elle marque un contraste avec les approches contrefactuelles, notamment, qui visent à identifier « ce qui marche » indépendamment du contexte de mise en place. Mais ce qui distingue l'analyse de contribution d'autres approches, c'est qu'elle rejette aussi l'idée que le rôle de l'évaluation soit d'établir un impact de façon irréfutable : dans un contexte complexe, son objectif n'est pas de prouver les effets des interventions, mais bien de réduire l'incertitude quant à leur contribution aux changements observés. C'est en effet cette incertitude qui est ici considérée comme préjudiciable à la décision et, plus globalement, à la fabrique de l'action publique.

Construction de la théorie (*theory-building*)

Tout le processus de l'analyse de contribution consiste ainsi à réduire progressivement l'incertitude sur les effets de l'intervention évaluée. La première phase de construction de la théorie consiste, comme pour toute EBT, à poser une question relative aux liens de cause à effet que l'on veut investiguer et à élaborer des hypothèses causales en réponse à cette question. Cette dernière porte le plus souvent sur des contributions de l'intervention à des changements recherchés. *Imaginons un plan interministériel visant à prévenir ou prendre en charge les violences sexuelles dans des établissements d'enseignement supérieur. La question*

posée pourra consister à demander « En quoi le plan engagé a-t-il contribué à la réduction effective des violences sexuelles et à la meilleure prise en charge de leurs suites? ».

Le niveau des violences et les réponses apportées dans les établissements, cependant, sont des changements de société, qui ne dépendent que très partiellement d'un quelconque plan ministériel. De fait, dans l'analyse de contribution, on ne préjuge pas que ces changements sont dus à l'intervention. Au contraire, on part du principe que tout changement est le résultat d'une multitude de causes entrelacées, dont (peut-être) l'intervention. C'est ainsi que l'analyse de contribution part du changement (ici, *l'évolution des violences sexuelles*) pour rechercher des contributions, plutôt que de l'intervention évaluée (*le plan interministériel*).

Ce à quoi s'attache donc l'analyse de contribution, dans cette phase initiale, c'est à expliciter ce en quoi pourrait consister la contribution de l'intervention (parmi d'autres facteurs) et à s'assurer qu'une telle contribution est plausible. Par plausible, on entend que cette contribution, sans être vérifiée, est néanmoins vraisemblable : elle pourrait avoir lieu dans le contexte de l'intervention évaluée.

Plus le cadre de l'intervention est complexe, et plus ce travail d'enquête initiale peut prendre du temps. La plausibilité d'une hypothèse n'est en effet pas jugée *in abstracto* : elle est estimée au regard de la convergence entre les observations, expériences et opinions informées des parties prenantes, de sa proximité avec des hypothèses validées dans d'autres cadres présentant des similarités avec l'intervention évaluée, sur la possible significativité de l'intervention par rapport à d'autres facteurs, sur de premiers indices d'un effet possible, etc.

Cette phase s'appuie généralement sur une première collecte de données empiriques (des échanges avec les parties prenantes, une analyse documentaire ou une revue de littérature) qui permet d'aboutir à des « hypothèses de contribution » (*contribution claims*, littéralement

« allégations relatives à la contribution ») ainsi qu'à des explications alternatives (c'est-à-dire des affirmations portant sur d'autres facteurs pouvant plausiblement expliquer les changements observés). Dans notre cas, une évaluation étudierait les changements intervenus (ou non) dans les violences sexuelles et dans les pratiques des établissements au cours des dernières années pour identifier des contribution claims. Si un certain nombre d'établissements ont changé drastiquement leurs pratiques sur ce domaine, c'est peut-être que le plan incluait une obligation de mettre en place des stratégies de lutte contre les violences sexuelles et de rendre compte des progrès annuellement; mais aussi peut-être que les acteurs déjà en faveur de démarches actives contre ces violences dans l'administration se sont servis de ce plan pour soutenir leur agenda interne; ou encore que des groupes étudiants s'en sont servis pour faire plier des administrations réticentes. Chacune de ces trois hypothèses, si elle s'appuie sur des exemples, un cadre théorique convaincant, etc. peut devenir une hypothèse de contribution.

À ce stade, le niveau d'incertitude quant aux effets de l'intervention est donc déjà réduit par rapport à la situation initiale : certaines affirmations ont été rejetées, d'autres apparaissent plus ou moins plausibles en l'état d'avancement de l'évaluation. Celles qui sont retenues sont étudiées dans l'étape suivante.

Mise à l'épreuve de la théorie (theory-testing)

Seules les hypothèses suffisamment plausibles (ou celles jugées particulièrement importantes pour les parties prenantes) sont en effet testées en profondeur. Dans l'analyse de contribution, il est possible d'utiliser une très vaste palette d'outils ou de méthodes, qualitatives ou quantitatives, pour estimer les changements et tester à charge et à décharge les hypothèses de contribution, en combinaison avec d'autres facteurs. Lors de ce processus, les hypothèses de contribution ne sont pas

validées ou écartées. Elles sont plutôt progressivement étoffées, passant par exemple de « l'intervention contribue de telle façon » à « lorsque les conditions x et y sont réunies, l'intervention contribue de telle façon, sauf si un évènement z apparaît », aboutissant ainsi à des « paquets causaux » (*causal packages*) réunissant plusieurs facteurs associés aux changements observés. L'analyse de contribution peut également s'attacher à identifier les chemins d'impact et les mécanismes sous-jacents qui expliquent ces contributions. *Par exemple, dans notre cas, peut-être que la phase de mise à l'épreuve permettrait de montrer que la mise à l'agenda du dialogue de gestion entre tel ministère et les établissements de la question des violences sexuelles a eu des conséquences directes en termes de mise en place d'un dispositif de remontée des violences; mais que tous les ministères n'ont pas réellement saisi cette question dans leur dialogue de gestion. Idéalement, la suite de la collecte consisterait alors à vérifier si des dispositifs de remontée des violences existent dans les établissements dépendant des autres ministères, et pourquoi.*

L'analyse de contribution ne préconise pas d'approche particulière pour établir l'inférence causale. Une démarche possible consiste à identifier, à l'image du « traçage de processus », une série de tests empiriques. Ces tests définissent chacun une condition qui doit être satisfaite pour qu'on puisse conclure que l'intervention contribue bien aux changements observés. Des tests peuvent également porter sur d'autres facteurs pouvant, de façon plausible, expliquer les changements. Tous les outils de l'évaluation et, plus largement, des sciences sociales, qualitatifs ou quantitatifs, peuvent être employés pour mener ces tests : entretiens, études de cas, analyses documentaires, mais aussi enquêtes, analyses statistiques... peuvent être mobilisés. La combinaison de ces outils permet, par triangulation, de renforcer (ou de réduire) le degré de confiance dans la contribution et d'aboutir aux constats et aux conclusions de l'évaluation. L'évaluation réaliste peut également être mobilisée ici pour identifier des mécanismes sous-jacents aux relations causales.

Une dernière spécificité de l'analyse de contribution est d'aboutir à des récits de contribution (*contribution story*). Le récit de contribution réunit au départ les *contribution claims*, progressivement rendues plus robustes par la collecte et l'analyse. Il a vocation à consolider, à compléter ou à défier les récits dominants qui sous-tendent l'intervention évaluée. Contrairement à une évaluation contrefactuelle par exemple, qui cherche à convaincre par la quantification, l'analyse de contribution s'appuie ainsi sur des narratifs appuyés par des éléments de preuve, qui peuvent ensuite être utilisés dans la fabrique de l'action publique. Dans notre cas, peut-être que le récit de contribution ferait apparaître la façon dont les parties prenantes déjà engagées dans la lutte contre les violences sexuelles se sont saisies du plan interministériel pour faire pencher la balance en leur faveur dans la gouvernance interne des établissements, au détriment d'un narratif national basé sur le contrôle par l'État des pratiques des établissements.

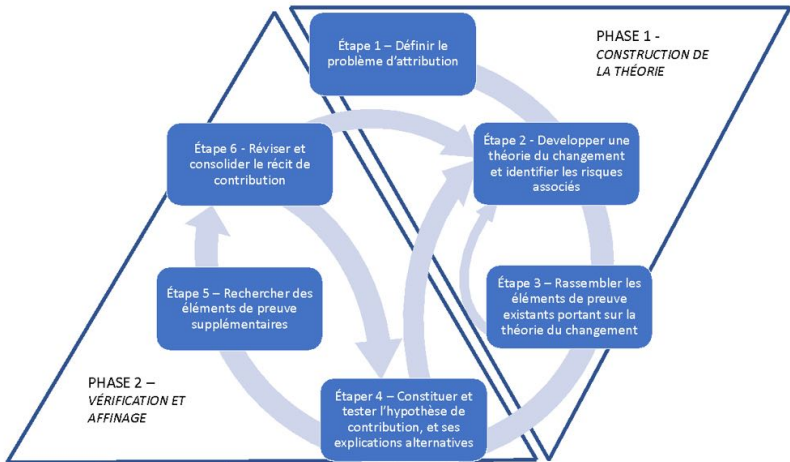


Figure 1 : Un processus en deux phases (Ton, Giel, 2021. « Development Policy and Impact Evaluation: Learning and Accountability in Private Sector Development ». In Handbook of Development Policy, par Habib Zafarullah et Ahmed Huque, 378-90. Edward Elgar Publishing, p. 380. Traduction de l'auteur.

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

L'analyse de contribution est essentiellement utilisée *ex-post*, bien qu'il existe des tentatives pour l'utiliser *chemin faisant*. Elle est particulièrement indiquée pour les cas où la contribution d'une intervention aux changements attendus est très incertaine, ou semble improbable, mais où cette contribution revêt un intérêt stratégique pour les parties prenantes : par exemple parce que les attentes sont très fortes vis-à-vis de cette contribution, ou parce que de cette contribution dépend la poursuite ou non de l'intervention.

Le travail réalisé dans la phase de construction de la théorie, parce qu'il permet de formuler des contributions plausibles qui s'écartent souvent des objectifs affichés, est particulièrement utile pour la gestion stratégique ou la reconception des interventions.

L'analyse de contribution se prête particulièrement bien à des démarches collaboratives ou participatives, permettant aux parties prenantes d'échanger sur les hypothèses de contribution et sur les conditions dans lesquelles elles ont des chances de se vérifier ou non. Les récits de contribution qu'elle produit, s'ils sont débattus et appropriés par les parties prenantes, fournissent une base utile à des réorientations stratégiques. Dans leur forme finale, les hypothèses de contribution, parce qu'elles sont explicatives et contextualisées, sont également utiles pour faire évoluer l'intervention ou les pratiques des acteurs et actrices impliquée-e-s.

III. L'exemple de la contribution d'une Fondation à la recherche dans les sciences du vivant

Une Fondation soutient sur le long terme (financement, accompagnement) des équipes et des institutions de recherche de haut niveau dans le domaine des Sciences du vivant¹. Les responsables de la Fondation sont conscient·e·s que les résultats des travaux financés ne peuvent pas être attribués seulement à leur soutien : les équipes de recherche sont en effet le principal moteur des résultats obtenus; elles s'appuient généralement sur une pluralité de financements; elles s'inscrivent dans des tendances de la recherche, font suite à des recherches passées et travaillent en lien avec d'autres équipes dans le monde. Enfin, les apports que peut avoir la Fondation sont indissociables du contexte de la recherche (sous-financement de la recherche en France, concurrence internationale...). Néanmoins, ses responsables pensent tout de même que son concours peut être significatif, et ils souhaitent l'explorer.

Eu égard à la diversité des projets soutenus (soutien individuel ou collectif, travaux de recherche, équipements, démarches pluridisciplinaires...), plusieurs théories du changement sont initialement conçues et alimentées par une collecte exploratoire (analyse documentaire, entretiens). Cette phase initiale aboutit à une première ébauche de « méta-théorie » du changement (réunissant les différentes théories élaborées), dans laquelle un certain nombre d'hypothèses de contribution (*contribution claims*) sont proposées. Celles-ci diffèrent en particulier selon la maturité du projet soutenu, et du type de soutien. Pour chacune de ces hypothèses de contribution, des tests empiriques sont élaborés, de façon à estimer le degré de confiance qu'il est possible

1. Cet exemple tiré d'une évaluation réelle a été simplifié à des fins pédagogiques.

d'avoir dans la réalité de ces contributions. Ces hypothèses sont ensuite soumises à examen à travers les tests afférents dans une série d'études de cas portant sur des projets soutenus par la Fondation.

L'analyse croisée des études de cas permet d'affiner et de circonscrire les contributions de la Fondation aux projets qu'elle soutient. Au total, 8 contributions principales sont identifiées, passant par différents chemins : par exemple, un financement de la Fondation peut contribuer à la pérennité d'un projet par son engagement sur la durée, mais aussi parce qu'il apporte de la légitimité au projet qui peut alors capter d'autres financements. La Fondation n'active pas toujours ces 8 contributions, mais son apport est plus important lorsque plusieurs sont activées sur un même projet. Le récit de contribution insiste sur l'inscription de ces contributions dans des facteurs explicatifs communs : par exemple le choix pertinent de chercheurs et de chercheuses qui savent utiliser les financements supplémentaires pour aller plus loin, ou tester ce qu'ils et elles n'auraient pu tester autrement; ou la relation de confiance mise en place, avec une grande liberté donnée aux équipes de recherche (qui se traduit notamment par des attentes minimales en termes de rendu compte des financements). Cette dimension humaine est aussi ce qui explique que ses apports soient plus marquants dans le soutien aux équipes de recherche, plutôt qu'aux institutions. L'évaluation alimente ainsi les évolutions stratégiques de la Fondation en identifiant les situations dans lesquelles son apport peut être le plus important et les choix qu'une réorientation supposerait en termes de moyens humains et financiers.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

La qualité de l'analyse de contribution se juge essentiellement à la capacité à travailler dans un *continuum* de plausibilité, ce qui signifie pouvoir considérer au départ un certain nombre d'hypothèses relatives aux facteurs qui sous-tendent les changements observés, dont l'intervention, à les passer en revue, identifier les plus plausibles, pour ensuite les tester et les étoffer progressivement.

Ces dernières années, le terme d'analyse de contribution est parfois utilisé comme synonyme apparemment flatteur d'évaluation basée sur la théorie. Parmi les principaux critères permettant de les différencier, on peut notamment citer :

1. la démarche itérative (dite *abductive*) de l'analyse de contribution (les hypothèses sont constamment révisées tout au long de l'évaluation);
2. le fait que la recherche des contributions démarre par les changements attendus et remonte à rebours vers l'intervention, plutôt que l'inverse;
3. une collecte d'information visant à progressivement contextualiser et étoffer les hypothèses de contribution;
4. le soin apporté à tester des explications alternatives et
5. la dimension narrative des résultats, sous la forme d'un récit de contribution.

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

L'analyse de contribution apporte des constats crédibles et utiles pour la fabrique de l'action publique dans des situations très particulières, qui semblent initialement très compliquées à évaluer. Elle doit sa crédibilité à son processus itératif, qui peut être rendu transparent dans une démarche participative. Le fait que les parties prenantes soient associées à chacune des étapes et la traçabilité des tests opérés, de même que l'humilité affichée de la démarche amènent un grand degré de confiance, base essentielle à l'utilisation des résultats.

Néanmoins, il faut garder en tête que le processus de l'analyse de contribution est lui-même incertain : on ne sait pas au départ quelles hypothèses de contribution seront testées et comment. Il est généralement nécessaire d'ouvrir la focale en début d'évaluation pour comprendre le contexte dans lequel se situe l'intervention, quelles interventions ou facteurs expliquent les changements constatés. Cette phase initiale, qui va consister à décrire les changements constatés, est ce qui fait tout l'intérêt de l'analyse de contribution par rapport à d'autres approches qui saisissent les interventions de façon presque « hors sol ». Mais cette phase peut être extrêmement chronophage, d'autant qu'elle dépend essentiellement de sources secondaires, externes à l'évaluation, et il faut trouver le bon niveau d'épaisseur à donner à la description – l'évaluation n'ayant pas vocation à être exhaustive.

Comme pour toute EBT, il faut prendre garde à ne pas surestimer les contributions, même si le fait de partir des changements plutôt que de l'intervention elle-même réduit ce risque. Une solution est l'application systématique de tests empiriques portant sur l'intervention et sur les explications alternatives. Mais cette solution peut également s'avérer très lourde et source de confusion, notamment lorsque les tests sont trop

nombreux ou mal calibrés (c'est-à-dire qu'ils ne permettent pas suffisamment de faire varier le degré de confiance dans une hypothèse de contribution).

À noter que là où l'évaluation réaliste et le traçage de processus sont plutôt utilisés à l'échelle projet ou pour tester un seul chemin d'impact, l'analyse de contribution est plutôt utilisée à l'échelle de programmes ou de politiques publiques, lorsque les acteurs impliqués et les chemins d'impact sont nombreux. Cette visée plus large est ce qui fait l'intérêt de l'analyse de contribution, mais elle renforce les incertitudes décrites ci-dessus.

Quelques références bibliographiques pour aller plus loin

L'analyse de contribution est née sous la plume de John Mayne au tournant des années 2000. On pourra lire les deux articles suivants, le premier marquant le début de la prise en compte de la complexité par l'analyse de contribution et le second présentant un état des débats et des évolutions de l'analyse de contribution en 2019 :

Mayne, John. 2012. Contribution analysis: Coming of age?. *Evaluation* 18 (3): 270-80. <https://doi.org/10.1177%2F1356389012451663>

Mayne, John. 2019. Revisiting Contribution Analysis. *Canadian Journal of Program Evaluation*. 34 (2). <https://doi.org/10.3138/cjpe.68004>

Les articles suivants témoignent de l'opérationnalisation progressive de l'approche dans les années 2010. Le premier rend compte d'un certain nombre d'obstacles pratiques et des façons dont des praticien-ne-s peuvent les dépasser; le second est un exemple emblématique d'une situation dans laquelle l'intervention évaluée n'est clairement pas le

principal moteur des changements attendus; le troisième donne un exemple de l'usage de l'analyse de contribution dans le développement du secteur privé :

Delahais, Thomas. et Toulemonde, Jacques. 2012. Applying Contribution Analysis: Lessons from Five Years of Practice. *Evaluation* 18 (3): 281-93. <https://doi.org/10.1177/1356389012450810>

Delahais, Thomas. et Toulemonde, Jacques. 2017. Making Rigorous Causal Claims in a Real-Life Context: Has Research Contributed to Sustainable Forest Management? *Evaluation* 23 (4): 370-88. <https://doi.org/10.1177/1356389017733211>

Ton, Giel. 2021. Development Policy and Impact Evaluation: Learning and Accountability in Private Sector Development. In *Handbook of Development Policy* par Zafarullah, Habib. et Huque, Ahmed. 378-90. Edward Elgar Publishing. <https://doi.org/10.4337/9781839100871.00042>

23. Récolte d'incidences

GENOWEFA BLUNDO CANTO

Résumé

La récolte d'incidences est une approche qualitative en évaluation *ex post*. Plutôt que de tester un impact spécifique de l'intervention étudiée, elle consiste, en collaboration avec les parties prenantes, à identifier et collecter des preuves empiriques concernant des effets intentionnels ou non, à étudier comment ces effets ont été produits et si et comment l'intervention a pu jouer un rôle dans ce processus, avant de valider ces résultats à l'aide de sources externes. La récolte d'incidences est une approche axée sur l'utilisation : elle vise à produire des connaissances pour l'action. Elle est particulièrement utile dans le cas d'interventions complexes, quand les effets d'une intervention ne sont pas connus ou identifiés au préalable ou dans le cas d'une intervention significativement modifiée depuis son lancement.

Mots-clés : Méthodes qualitatives, étude de cas, changements observables, énoncé des incidences, récolte, impact

I. En quoi consiste cette approche?

La récolte d'incidences est une approche qualitative en évaluation dans laquelle les personnes qui assurent la récolte (les évaluateurs et évaluatrices), identifient, formulent, vérifient, analysent et interprètent des effets (incidences). Les incidences sont définies comme des changements observables dans le comportement d'individus ou de collectifs tels que des groupes, des communautés, des organisations, des

institutions. Les incidences sont des changements dans les pratiques, les relations, les politiques, les actions et les activités de ces individus ou collectifs. Cela peut être l'augmentation du nombre de femmes inscrites à une formation professionnelle ou l'utilisation d'une pratique de gestion innovante par une organisation d'agriculteurs et d'agricultrices. Ces incidences peuvent être positives ou négatives, intentionnelles ou non, attendues ou inattendues. Elles ne doivent pas être confondues avec les impacts, c'est à dire les conséquences ultimes de l'intervention sur des dimensions du bien-être des bénéficiaires ciblé-e-s, comme la santé ou l'éducation. Selon le lexique de la récolte d'incidences, les impacts (par exemple, en matière de santé ou d'éducation) ne peuvent être obtenus sans un changement de comportement (par exemple, l'application de pratiques alimentaires, la participation à des formations). Ce sont donc ces comportements qui sont ciblés par la récolte d'incidences.

La récolte d'incidences fournit des preuves empiriques sur les incidences (*outcomes*) obtenues et la façon dont une intervention y a contribué, ainsi que sur la signification de ces résultats à la lumière des questions de l'évaluation. La récolte d'incidences ne se concentre pas sur le progrès ou la réalisation des résultats prévus, attendus ou planifiés de l'intervention, mais recueille des preuves sur ce qui a été réalisé et travaille à rebours pour identifier si et comment l'intervention a contribué à ces changements. Dans la récolte, tous les changements qui se sont réellement produits sont collectés, ce qui permet de saisir également les résultats involontaires, positifs et négatifs. L'évaluateur ou l'évaluatrice facilite leur identification et la recherche de preuves de la manière dont ils ont été obtenus et de la contribution de l'intervention. La récolte d'incidences est axée sur l'utilisation : son but est de servir les utilisations des résultats de l'évaluation par les utilisateurs prévus, c'est-à-dire ceux qui prendront des décisions sur la base de ces résultats. L'accent est mis sur l'apprentissage à partir de l'évaluation afin d'agir à partir de ces apprentissages.

L'élément clé de la récolte d'incidences est l'énoncé des incidences (*outcome statement*) qui décrit le changement, qui l'a fait, quand et où, quelle était la contribution plausible des activités, stratégies et produits de l'intervention à chaque changement, et l'importance du changement. À titre d'exemple : depuis 2015, le département de l'agriculture du Sénégal publie un bulletin d'information présentant des prévisions météorologiques et des recommandations agricoles ciblées en utilisant un langage et un format adaptés aux agriculteur·rice·s et au grand public, démocratisant ainsi l'accès aux connaissances scientifiques. L'intervention a contribué à ce changement grâce à une étude sur les canaux et les formats de communication préférés des agriculteur·rice·s et à une série de formations en communication scientifique.

La récolte d'incidences se fait en six étapes, à travers lesquelles les énoncés des incidences sont affinés et démontrés empiriquement. La première étape consiste à concevoir la récolte afin de répondre aux utilisations prévues des résultats, telles que définies par leurs utilisateurs et utilisatrices primaires. La deuxième étape consiste à examiner la documentation pour identifier et formuler des projets d'énoncés de résultats. La troisième étape consiste à ouvrir un dialogue avec les personnes qui connaissant la contribution de l'intervention à ces changements, lesquelles examinent et affinent les énoncés d'incidences avec l'équipe d'évaluation jusqu'à ce qu'un ensemble d'énoncés précis soit identifié. Définir clairement ce qui a changé, la contribution de l'intervention et l'importance du changement permet de délimiter les incidences qui sont prises en compte dans l'évaluation et celles qui ne le sont pas. L'idée est de s'efforcer d'obtenir un nombre réduit de changements vérifiables pour lesquels il faudra recueillir des preuves empiriques à l'étape suivante. La quatrième étape est la démonstration empirique des incidences par des sources externes indépendantes de l'intervention mais connaissant les changements étudiés et pouvant valider la contribution de l'intervention. La corroboration permet de vérifier l'exactitude des incidences, mais aussi d'enrichir la compréhension du changement et de la contribution d'autres acteurs

ou interventions. D'autres changements liés à l'intervention que les utilisateurs et utilisatrices primaires n'avaient pas identifiés peuvent émerger lors de l'étape de corroboration. La cinquième étape consiste à analyser et interpréter les énoncés d'incidences, en systématisant les preuves empiriques pour répondre aux questions d'évaluation définies lors de l'étape de conception. La sixième étape consiste à aider les utilisateurs et utilisatrices à utiliser les résultats de l'évaluation pour les usages prévus.

Les 6 étapes de la récolte d'incidences

1. Conception de la récolte
2. Examen de la documentation, première rédaction des énoncés d'incidences
3. Dialogue avec les parties prenantes
4. Démonstration empirique des incidences à l'aide de sources externes
5. Analyse et interprétation des énoncés d'incidences
6. Soutien à l'utilisation des résultats de la recherche par les parties prenantes

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

La récolte d'incidences est une méthode d'évaluation rétrospective ou *ex post*, mais elle peut également être utilisée pour suivre les progrès pendant une intervention. Elle est particulièrement utile pour les interventions complexes où le contexte est imprévisible, incertain, dynamique, et où les actions prévues sont susceptibles de changer. La récolte d'incidences est particulièrement appropriée pour : 1) suivre et

évaluer des interventions sur des défis émergents pour lesquels peu d'informations et de preuves existent; 2) générer des preuves empiriques des changements obtenus par une intervention qui n'a pas prédéfini de changements ou qui avait des changements prédéfinis très généraux; 3) fournir des preuves sur les changements générés directement et indirectement par une intervention, intentionnels ou non; 4) évaluer une intervention qui a changé de manière significative par rapport à ce qui était initialement prévu.

Cette approche peut être utilisée pour suivre une intervention en cours en collectant périodiquement et systématiquement des informations et des enseignements sur les changements sociaux obtenus et la contribution de l'intervention à ceux-ci. Elle peut être combinée avec des méthodes qualitatives ou quantitatives qui répondent à d'autres questions d'évaluation. Par exemple, elle peut compléter les résultats des méthodes qui quantifient les impacts liés au bien-être auxquels ont conduit les changements récoltés.

La récolte d'incidences se concentre sur la contribution des actions de l'intervention au changement social, c'est-à-dire sur la relation plausible et logique entre l'intervention et le changement, plutôt que sur la part exacte attribuable à cette action. Contrairement à d'autres approches basées sur la contribution (voir chapitre séparé sur l'analyse de contribution), la récolte d'incidences ne démarre pas par l'identification de la relation cause-effet à des changements prédéfinis et intentionnels. Au contraire, elle récolte « tous » les résultats observés et travaille ensuite à rebours pour reconstruire la contribution de l'intervention à ces changements. En effet, l'intérêt principal de la récolte d'incidences réside dans sa capacité de rendre compte de dynamiques de changement social en adoptant une approche ouverte et large dans l'identification de ces changements, même lorsque celles-ci sont inattendues ou involontaires. Pour cette raison, elle est particulièrement adaptée à l'évaluation d'interventions complexes où l'incertitude, les contextes dynamiques et l'adaptation sont fréquents.

III. Un exemple de l'utilisation de cette approche : l'extension d'une politique d'information météorologique et climatique au Sénégal

Ce qui suit présente une étude de cas utilisant la récolte d'incidences dans le cadre d'une approche d'évaluation plus large pour évaluer les résultats de l'extension des services d'information météorologique et climatique (SMC) au Sénégal et comment des actions de recherche ont contribué à ces résultats (Blundo-Canto et al. 2021). Les SMC impliquent la production, la traduction, la transformation, la transmission, l'accès et l'utilisation d'informations scientifiques sur la météo et le climat pour soutenir la prise de décision. Au Sénégal, la diffusion de prévisions météorologiques et climatiques accompagnées de recommandations pour les secteurs et les acteurs économiques s'est étendue, passant de projets de recherche pilotes à une stratégie de niveau national, en deux décennies. L'évaluation des résultats de cette généralisation des services s'inscrivait dans une double logique de reddition de comptes et d'apprentissage, et reposait sur trois composantes. La reconstruction de l'histoire de l'innovation (Douthwaite et Ashby 2005) : la chronologie détaillée et l'interconnexion des événements, des facteurs, des actions et des individus et collectifs qui ont marqué la mise à l'échelle des SMC. La récolte d'incidences (Wilson-Grau 2018) : l'évaluation des changements dans les pratiques, les relations, les politiques, les actions et les activités observables des individus et collectifs impliqué-e-s et affecté-e-s par la mise à l'échelle des SMC, et la contribution des partenariats de recherche à ces changements. L'analyse du chemin de l'impact (Douthwaite et al. 2003) : la chaîne de causalité menant des actions de recherche de l'Agence nationale de l'aviation civile et de la météorologie (ANACIM) et de ses partenaires aux résultats identifiés et à leurs effets perçus à moyen et long terme.

Les six étapes de l'approche de récolte d'incidences ont guidé la mise en œuvre des trois composantes. Dans l'étape de conception, l'équipe d'évaluation a discuté avec l'ANACIM de la conception de la récolte, en identifiant les sources documentaires et les acteurs et actrices clés qui devraient participer à l'étape de formulation. Lors de la deuxième étape, la documentation existante a été examinée afin de reconstituer l'histoire de l'innovation et de pré-identifier les résultats qui pourraient être liés au processus de mise à l'échelle, qui ont été discutés avec l'Agence. Lors de l'étape de formulation, un atelier réunissant 16 représentant·e·s d'acteur·rice·s nationaux et locaux impliqués dans le processus de mise à l'échelle a été organisé afin de reconstituer les principaux événements, individus et collectifs, actions et facteurs contextuels. L'atelier a permis d'identifier d'autres individus et collectifs impliqué·e·s dans le processus et certaines incidences supplémentaires. Les changements identifiés au cours des trois étapes précédentes ont été systématisés et formulés, puis étayés par des entretiens individuels avec 44 informateurs et informatrices compétent·e·s et indépendant·e·s de l'acteur principal du changement, l'ANACIM. Néanmoins, la particularité du processus politique étudié, qui consiste à passer d'un projet pilote à une action d'envergure nationale impliquant de nombreux partenaires et secteurs, a fait que certain·e·s des participant·e·s à l'atelier ont été interrogé·e·s dans le cadre du processus de démonstration empirique afin de valider et de fournir des preuves des résultats pour lesquels ils avaient par ailleurs été sollicités en tant qu'informateurs et informatrices compétent·e·s. Grâce aux preuves fournies par l'étape de validation, les résultats ont été affinés, de même que la contribution des actions de recherche de l'ANACIM et de ses partenaires, ainsi que d'autres acteur·rice·s et facteurs. Afin de soutenir l'utilisation, trois ateliers de restitution au niveau national et local ont été réalisés, dans lesquels les résultats de la récolte d'incidences ont été présentés et discutés par les participant·e·s, ainsi que les actions possibles sur la base des connaissances générées.

Les principales conclusions de l'approche combinant la récolte d'incidences, les histoires d'innovation et l'analyse des chemins d'impact peuvent être résumées comme suit. Au cours des deux dernières décennies, les services météorologiques et climatiques ont servi d'instruments politiques clés pour faire face à la variabilité accrue des précipitations et les événements climatiques extrêmes qui affectent les communautés rurales vulnérables du Sahel ouest-africain. L'histoire de l'innovation commence dans les années 1980 lorsque, suite à une sécheresse dévastatrice, le centre régional d'agriculture, d'hydrologie et de météorologie créait le premier groupe de travail multidisciplinaire pour faciliter le développement des services météorologiques et climatiques, leur interprétation, leur diffusion et leur utilisation. Dans les années 2000, l'Agence nationale de météorologie du Sénégal s'associait à des acteurs de la recherche nationale et internationale pour mettre en place des groupes de travail multidisciplinaires décentralisés pilotes, afin de faciliter l'assimilation des prévisions et des recommandations au niveau local. Les informations climatiques étaient combinées avec des conseils agricoles dans un langage qui utilisait les concepts, les habitudes et les pratiques locales pour rendre ces informations exploitables par les agriculteurs et agricultrices. Les groupes de travail multidisciplinaires se réunissaient régulièrement pendant la saison des pluies pour discuter de la transmission des prévisions et des recommandations. Des personnes reconnues dans leurs communautés agricoles étaient formées aux concepts de prévision et à leur interprétation afin d'accroître l'assimilation par les autres agriculteurs. Au cours des années menant à 2018, les SMC et les groupes de travail multidisciplinaires ont été séquentiellement étendus à la plupart des départements du Sénégal.

La récolte d'incidences nous a permis d'identifier comment les informations climatiques étaient intégrées dans les plans, stratégies et programmes d'adaptation sectoriels et nationaux, ainsi que dans la coordination des actions de multiples acteurs et actrices au niveau local. Elle nous a également permis d'identifier d'autres secteurs que l'agriculture, notamment la pêche, l'énergie et la protection des

ressources en eau, qui utilisaient les SMC, montrant ainsi que les résultats générés dépassaient les frontières institutionnelles, sectorielles et de gouvernance. Au-delà des actions de l'ANACIM et de ses partenaires, ce processus a été soutenu par un environnement de financement mondial favorable. En combinant la récolte d'incidences avec l'analyse des chemins d'impact, il est apparu que l'extension des SMC à de nouveaux et nouvelles utilisateurs et utilisatrices, secteurs et usages, s'est faite selon cinq axes : 1) l'amélioration continue des SMC, 2) l'émergence et la consolidation des facilitateurs de SMC, 3) l'inclusion des SMC dans la planification des actions, 4) la mobilisation active pour soutenir la mise à l'échelle des SMC, et 5) l'habilitation des acteurs et actrices à assumer de nouveaux rôles. Les facteurs sous-jacents au processus de mise à l'échelle peuvent être résumés comme des actions intentionnelles des partenaires de recherche, y compris le renforcement des capacités, le partage des connaissances, les plateformes d'action, et la création d'opportunités d'interaction; le soutien financier national et international; et un environnement politique favorable. L'amélioration continue des SMC grâce à la prise en compte des retours de ses utilisateurs et utilisatrices a renforcé le processus de mise à l'échelle, ce qui a permis d'accroître l'accès de la population aux SMC. La récolte d'incidences a également permis de saisir les défis soulevés par l'expansion des SMC au fur et à mesure de l'apparition de nouveaux usages, utilisateurs et utilisatrices. En même temps, il existe une demande croissante pour des SMC de meilleure qualité et à grain plus fin, délivrés au bon moment pour prendre des décisions, ce qui nécessite des investissements importants. Les questions de confiance quant à la provenance de l'information, à la manière dont elle est produite et à la façon de l'interpréter peuvent constituer un obstacle au fur et à mesure que les SMC touchent de plus en plus d'utilisateurs et utilisatrices, si des campagnes de renforcement des capacités ne sont pas mis en place. Les partenariats public-privé pourraient jouer un rôle important, mais à l'heure actuelle, la participation du secteur privé à la diffusion des SMC est limitée. Les résultats de l'évaluation ont été utilisés dans une logique de reddition de comptes par les partenaires de recherche impliqués, mais aussi pour la production

de connaissances scientifiques sur la mise à l'échelle de ces instruments politiques, et pour discuter des principaux défis que les individus et collectifs impliqués dans la production, la diffusion et l'utilisation des SMC doivent surmonter.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

La récolte d'incidences se concentre sur l'évaluation des résultats, et non des impacts. Son objet n'est pas de compter des bénéficiaires ou de mesurer les effets qu'ils et elles ressentent. Il est important de le préciser pour juger de la qualité de cette méthode. Son objectif est d'identifier les changements intentionnels, non intentionnels, attendus ou inattendus dans les pratiques, les relations, les politiques, les actions ou les activités des individus et collectifs et d'évaluer la contribution d'une intervention à ces changements. Une telle évaluation se concentre sur l'utilisation des résultats par les utilisateurs et utilisatrices prévu·e·s, pour les usages prévus. Par conséquent, sa qualité doit être jugée en fonction de ces usages prévus (par exemple, la reddition de comptes ou l'apprentissage pour une gestion adaptative). Par exemple, l'étape de démonstration empirique est importante lorsque l'objectif est la reddition de comptes et le bilan d'ensemble d'une politique, mais elle peut être négligée ou effectuée de manière plus légère lorsque l'objectif est l'apprentissage interne ou le suivi.

Pour juger les énoncés d'incidences, la récolte d'incidences utilise les critères SMART : chaque déclaration doit être Spécifique, suffisamment détaillée pour être appréciée par n'importe quel lecteur ou lectrice; Mesurable: fournissant des informations quantitatives et qualitatives vérifiables; Réalisée: un lien plausible entre la contribution de l'intervention et l'incidence peut être établi; Pertinente: l'incidence est significative à la lumière de l'objectif de l'intervention; opportune dans le

Temps: l'incidence s'est produite à proximité du moment où l'évaluation est réalisée, même si la contribution de l'intervention s'est produite un certain temps auparavant.

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

La caractéristique principale qui distingue la récolte d'incidences des autres méthodes d'évaluation est qu'elle se concentre sur les changements obtenus, qu'ils aient été planifiés ou non, ce qui permet de saisir les résultats involontaires ou inattendus, tant positifs que négatifs. La méthode fournit un moyen systématique et structuré d'identifier ces changements et de travailler à rebours pour déterminer si et comment l'intervention y a contribué. La récolte d'incidences mobilise des données quantitatives et qualitatives pour décrire les incidences. Cependant, elle ne fournit pas une évaluation quantitative de ces changements. Elle informe plutôt sur les processus et les stratégies qui ont conduit à un changement quantitatif, qui peut par ailleurs être mesuré par d'autres méthodes. Elle ne peut pas être utilisée pour mesurer l'impact. Lorsque la récolte d'incidences omet l'étape de justification empirique, elle est alors plus adaptée pour l'apprentissage interne que pour la reddition de comptes, car elle ne recourt alors pas à un processus de validation par des sources indépendantes et bien informées.

Comme d'autres approches axées sur l'utilisation, la récolte d'incidences se concentre sur l'utilité de l'évaluation pour ses utilisateurs et utilisatrices et pour les utilisations prévues. Celles-ci peuvent être de l'ordre de l'apprentissage, la prise de décision, la planification, la reddition de comptes, l'information des partenaires, etc., en fonction de ce qui a été convenu avec les utilisateurs et utilisatrices primaires au stade de la conception. Ce choix guide la manière dont la méthode est appliquée et le poids accordé à la justification des changements sociaux et de la

contribution de l'intervention. La mesure dans laquelle la contribution de l'intervention est évaluée dans la récolte sera plus élevée lorsque l'utilisation est la reddition de comptes, à la fin d'une intervention, que lorsque l'utilisation prévue est l'apprentissage pour favoriser une gestion adaptative pendant l'intervention.

Quelques références bibliographiques pour aller plus loin

Cette note méthodologique s'inspire largement de :

Wilson-Grau, Ricardo. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.

Blundo-Canto, Genowefa. et Andrieu, Nadine. et Soule Adam, Nawalyath. et Ndiaye, Ousmane. et Chiputwa, Brian. 2021. « Scaling Weather and Climate Services for Agriculture in Senegal: Evaluating Systemic but Overlooked Effects ». *Climate Services* 22 (Avril): 100216. <https://doi.org/10.1016/j.cliser.2021.100216>.

Références supplémentaires mentionnées :

Douthwaite, Boru. et Ashby, Jacqueline. 2005. « Innovation Histories: A Method for Learning from Experience ». *The Institutional Learning and Change (ILAC) Initiative*, 4. <https://hdl.handle.net/10568/70176>.

Douthwaite, Boru. et Kuby, Thomas. et van de Fliert, Elske. et Schulz, Steffen. 2003. « Impact pathway evaluation: an approach for achieving and attributing impact in complex systems ». *Agricultural Systems, Learning for the future: Innovative approaches to evaluating agricultural research*, 78(2): 243-65. [https://doi.org/10.1016/S0308-521X\(03\)00128-8](https://doi.org/10.1016/S0308-521X(03)00128-8).

24. Sécurisation culturelle

LOUBNA BELAID ET NEIL ANDERSSON

Résumé

Concept issu des sciences infirmières et ici appliqué de façon innovante à l'évaluation, la sécurisation culturelle désigne une démarche visant à faire en sorte que l'évaluation se déroule d'une façon « sûre » pour les parties prenantes et notamment les communautés minorisées cibles de l'intervention étudiée, c'est-à-dire que le processus d'évaluation évite de reproduire des mécanismes de domination (agression, déni d'identité...) liés à des inégalités structurelles. Pour cela, diverses techniques participatives sont mobilisées à toutes les étapes de l'évaluation. La sécurisation culturelle est compatible avec tous types de méthodes. Elle contribue à rendre l'évaluation plus pertinente et utile pour les parties prenantes, et est susceptible d'accroître leur autodétermination.

Mots-clés : Méthodes mixtes, participation, évaluation autochtone, évaluation attentive aux différences culturelles, inégalités, racisme, décolonialité, cartographie cognitive floue

I. En quoi consiste cette approche?

La démarche de sécurisation culturelle s'est développée en réponse au constat selon lequel le processus d'évaluation pouvait reproduire les mécanismes de domination liés à des inégalités structurelles, notamment vis-à-vis de peuples autochtones ou dans des contextes post-coloniaux. Par exemple, une recherche a évalué la perception de trois échelles

psychométriques utilisées pour diagnostiquer la dépression dans les populations Inuit et Mohawk du Québec. Les résultats de l'étude ont montré que les trois échelles n'étaient pas culturellement sûres. Les participants et participantes n'ont pas apprécié l'évaluation numérique, l'auto-évaluation, (par opposition à l'interaction de soutien) et l'accent mis sur les symptômes plutôt que sur les facteurs de soutien (Gomez Cardona et al. 2021).

La sécurisation culturelle vise à produire « un environnement pour les personnes où il n'y a pas d'agression, de contestation ou de dénis de leur identité, de qui ils sont, et de ce dont ils ont besoin » (Williams 1999). Une infirmière maorie a initialement développé le concept en réponse au racisme et à la discrimination à laquelle sont confrontés les Maoris dans les établissements de santé (Papps et Ramsden 1996). Une évaluation culturellement sûre implique que les parties prenantes sentent que leurs cultures sont respectées et renforcées par l'évaluation.

La sécurisation culturelle va au-delà d'une autre démarche plus couramment promue en évaluation à partir des notions de sensibilité et de compétence culturelles (on parle ainsi d'évaluation attentive aux différences culturelles ou *culturally responsive evaluation*). En effet, au-delà de la simple attention portée aux différences culturelles, la sécurisation culturelle prend en considération les déséquilibres de pouvoir, la discrimination institutionnelle, le racisme et les relations coloniales qui peuvent s'immiscer dans la conception et la mise en œuvre des services et des programmes (Curtis et al. 2019). Le concept se situe donc dans le spectre des théories critiques postcoloniales, et vise la justice sociale. Le concept de sécurisation culturelle a été étendu au-delà des communautés maories à tout groupe qui diffère de son prestataire de soins ou de service par l'âge, le genre, le statut socio-économique, l'origine ethnique, la religion ou le handicap (Smye et Browne 2002).

Le concept a attiré l'attention d'approches en recherche et en évaluation qui remettent en question les perspectives unidirectionnelles et conventionnelles centrées sur le point de vue de la personne qui mène

l'étude, avec une contribution ou des avantages minimales pour les participant-e-s (Smith 2012; Cram 2016; Katz et al. 2016). Il invite par ailleurs à revisiter les concepts, les méthodes, les valeurs et les approches d'évaluation issues des épistémologies occidentales, attirant l'attention sur la validité des visions du monde des groupes autochtones et minoritaires (Smith 2012; Belaid et al. 2022).

La sécurisation culturelle garantit que l'évaluation est bénéfique et pertinente pour ces communautés. Elle vise à renforcer leur pouvoir d'agir, et peut contribuer à accroître leur autodétermination (Gollan et Stacey 2021). La sécurisation culturelle modifie la direction de l'évaluation en intégrant la perspective des participant-e-s.

Cinq principes clés caractérisent la sécurisation culturelle en évaluation (Wilson et Neville 2009; Cameron et al. 2010; Andersson 2018) : (i) la participation (ii) le partenariat (iii) l'appropriation (iv) la réflexivité critique et (v) la protection des identités, valeurs, croyances, valeurs culturelles et vision du monde.

- (i) *La participation*: fait référence à l'implication des parties prenantes tout au long de l'évaluation (Cameron et al. 2010). La participation procédurale ou symbolique doit être distinguée de la participation authentique. La participation procédurale permet une contribution structurée des parties prenantes à des étapes spécifiques du processus, par exemple, des entretiens structurés avec des informateurs et informatrices clés. La participation symbolique peut impliquer un-e « représentant-e » des parties prenantes, rémunéré-e ou non, participant à certaines activités de l'évaluation. Une participation authentique comprend la co-appropriation de l'évaluation, un engagement actif dans l'analyse des données probantes et un rôle dans la conception des solutions issues des résultats de l'évaluation.

- (ii) *Le partenariat* formalise la participation, évoluant souvent vers une participation authentique au fur et à mesure que l'évaluation se déroule. Il s'agit d'établir des relations équitables entre l'équipe d'évaluation et les parties prenantes, qu'il s'agisse des communautés, des patient·e·s ou des employé·e·s. L'équipe d'évaluation doit clarifier dès le départ la portée potentielle de ces relations, s'efforcer de les maintenir et de permettre leur évolution au fur et à mesure que les parties prenantes augmentent leurs capacités tout au long de l'évaluation (Cameron et al. 2010).
- (iii) *L'appropriation du processus d'évaluation, des résultats et de la gouvernance*: la sécurisation culturelle permet aux parties prenantes de s'approprier le processus (Andersson 2018). Cela peut commencer par la circonscription de l'objet de l'évaluation, dans les limites des objectifs financés, ou encore le fait d'« avoir une voix » sur ce qui est évalué et comment, et le fait de s'impliquer dans les activités d'évaluation, y compris dans l'interprétation des résultats.
- (iv) *Réflexivité critique*: Le point de départ d'une évaluation culturellement sûre est que les évaluateur·rice·s réfléchissent à leurs valeurs et croyances, leur position sociale, leur pouvoir et leurs privilèges (Wilson et Neville 2009; Browne et al. 2016). Cela implique une prise de conscience de la relation historique entre les évaluateur·rice·s et les communautés autochtones et minoritaires, de l'histoire du colonialisme, du racisme systémique et de la discrimination à laquelle ces communautés pourraient encore être confrontées (Cameron et al. 2010).
- (v) *La protection* renforce l'éthique de la recherche en protégeant les groupes autochtones et minoritaires de l'exploitation et du renforcement des représentations ou explications négatives (Wilson et Neville 2009). Cela implique que leurs connaissances, leurs valeurs et leurs épistémologies soient également valorisées aux côtés des

épistémologies et des méthodes scientifiques occidentales (Cameron et al. 2010). Les communautés autochtones veulent ainsi voir l'évaluation enracinée dans leur vision du monde (Belaïd et al. 2022).

Plusieurs cadres et lignes directrices traitent de l'évaluation équitable avec les groupes autochtones et minoritaires (Wilson et Neville 2009; Cameron et al. 2010; Gollan et Stacey 2021). Nous présentons ici le cadre développé par Andersson et ses collègues pour préciser comment la sécurisation culturelle se déploie aux différentes étapes de l'évaluation (Cameron et al. 2010; Andersson 2018).

Formulation de l'objet de l'évaluation

Idéalement, une évaluation culturellement sûre émane d'une requête formulée par la communauté. Cela augmente la pertinence de l'évaluation, en s'assurant d'un alignement sur les priorités de la communauté. Dans la pratique, de nombreuses évaluations sont commanditées à partir d'un problème défini par les bailleurs de fonds, ce qui laisse moins de place pour formuler ou renommer l'objet de l'évaluation.

L'évaluation culturellement sûre choisit de miser sur les forces des communautés. Les groupes sociaux dominants qualifient souvent les groupes autochtones et minoritaires de « groupes à risque », « vulnérables » ou « marginalisés ». Ces appellations négatives ne reflètent pas nécessairement la façon dont ces communautés se perçoivent et réduisent considérablement l'espace et les conditions d'amélioration (Wilson et Neville 2009). Lorsque les parties prenantes définissent ou renomment l'objet de l'évaluation, en mettant en avant leurs forces pour l'aborder, cela réduit l'étiquetage négatif et ouvre la voie à l'amélioration.

La cartographie cognitive floue (CCF) permet à des groupes ou à des individus de formuler le problème d'évaluation en fonction de leur propre compréhension de celui-ci (Andersson et Silver 2019). Lors d'une session

de cartographie cognitive floue, les participant·e·s construisent un modèle causal flexible de la façon dont ils et elles voient le problème en fournissant les concepts et le lexique qui leur sont familiers. Ils et elles décrivent les facteurs qui influencent le problème et discutent des orientations et des forces de chaque relation ayant un impact sur le problème. Cette cartographie est appelée « floue » car elle évalue l'influence relative de chaque relation dans la carte : les participants et participantes sont ainsi invité·e·s à évaluer cette influence sur une échelle de 1 (faible) à 5 (fort). Cette démarche enrichit l'évaluation, mais modifie également son appropriation.

La CCF appuie la sécurisation culturelle de plusieurs façons. Ne nécessitant aucun niveau d'alphabétisation ou de langue, elle favorise l'équité et l'inclusion (Andersson et Silver 2019). Elle réduit la stigmatisation ou la honte que peuvent ressentir des individus ou des groupes qui font partie de groupes précédemment exclus. Lors des sessions de CCF, les groupes sont organisés par sexe, âge et type de parties prenantes (par exemple, patient·e·s, prestataires, responsables de programme) pour garantir la représentation de toutes les voix. Les facilitateur·rice·s rencontrent chaque groupe séparément, réduisant les déséquilibres de pouvoir entre les groupes. Idéalement, les facilitateur·rice·s devraient être du même sexe et du même âge, et partager la langue et les réalités socioculturelles des participant·e·s, afin de réduire la hiérarchie avec elles et eux. Les facilitateur·rice·s reçoivent une formation pour réduire les préjugés.

Les résultats de la carte cognitive floue peuvent éclairer l'évaluation au-delà de la formulation du problème. Elle permet une contribution des parties prenantes dans la conception du questionnaire, en combinant la littérature existante avec les opinions des parties prenantes et leur compréhension des mécanismes de changement (Dion et al. 2019; Dion et al. 2022). Giles et ses collaborateurs ont utilisé l'outil pour saisir comment une communauté mohawk comprend les facteurs qui influencent le diabète (Giles et al. 2007). Sarmiento et ses collègues ont utilisé l'outil pour explorer comment les communautés autochtones de l'État de

Guerrero perçoivent les facteurs qui influencent la santé maternelle, afin de mieux concevoir des interventions (Ivan Sarmiento, Paredes-Solís, et al. 2020; Iván Sarmiento, Zuluaga, et al. 2020).

Éthique

Les comités d'approbation éthique institutionnels et les comités d'éthique pour l'évaluation s'appuient presque invariablement sur les épistémologies occidentales, s'attendant à ce que tous les aspects de l'évaluation soient clarifiés avant le début de cette dernière (Cameron et al. 2010). Pourtant, la sécurisation culturelle implique la contribution des participant·e·s aux protocoles et aux outils d'évaluation, ce qui n'est généralement pas possible avant le début de l'évaluation. Les évaluations culturellement sûres devraient par ailleurs demander l'approbation des comités autochtones dans la mesure du possible et respecter les lignes directrices éthiques pour la recherche autochtone. De plus, elles devraient appliquer les principes d'appropriation, de contrôle, d'accès et de possession concernant la manière dont « les données et les informations des populations autochtones seront collectées, protégées, utilisées et partagées. Ces principes visent à accroître la gouvernance de l'information des populations autochtones » (Nations 1998).

Devis (design) de recherche

La sécurisation culturelle peut être appliquée aux évaluations qualitatives, quantitatives et mixtes. Andersson et ses collègues ont utilisé un essai randomisé contrôlé pour évaluer des interventions locales visant à réduire la violence domestique en partenariat avec 12 refuges pour femmes autochtones au Canada (Andersson et al. 2010). Ce sont les

directeurs/-trices des refuges qui ont requis l'essai contrôlé randomisé. Ils et elles ressentaient en effet la nécessité de montrer l'impact de leur programme afin de demander plus de financement (Andersson et al. 2010).

Développement des instruments et méthodes de collecte de données

Les équipes d'évaluation sont souvent encouragées à utiliser des questionnaires standardisés, pour bénéficier de leur validité et de leur fiabilité. De nombreux questionnaires conventionnels se concentrent sur les facteurs de risque, les déficits, plutôt que sur les forces et la résilience qui caractérisent de nombreuses visions du monde autochtones et de groupes minoritaires. La sécurisation culturelle exige plus de flexibilité, notamment dans les variables incluses dans les questionnaires. Il s'agit d'inclure les facteurs que les parties prenantes perçoivent comme importants, même s'ils ne font pas partie d'un questionnaire standardisé. Très souvent, il est possible de définir les thèmes d'un questionnaire par un processus participatif, par exemple, la cartographie cognitive floue précédemment mentionnée.

La collecte de données dans les évaluations culturellement sûres peut être quantitative ou qualitative. Chaque méthode peut introduire un biais substantiel lorsqu'elle est contrôlée de trop près par l'équipe d'évaluation. Le recrutement et la formation minutieuse d'une équipe locale de collecte de données, qui partage un contexte social et culturel similaire avec la communauté, peut être une meilleure stratégie. Non seulement cela augmente l'acceptabilité et les taux de réponse, mais l'évaluation favorise le développement de compétences dans la communauté.

Analyse des données et interprétations

Que ce soit à partir de méthodes qualitatives ou quantitatives, l'analyse et l'interprétation des données devraient refléter les visions du monde des Autochtones et des groupes minoritaires. Par exemple, dans le cadre de l'analyse thématique inductive, il existe des options de catégorisation et de codage participatives (Liebenberg, Jamal, et Ikeda 2020). Même lorsque l'analyse est informatisée, ce qui rend difficile une participation plus large, la vérification fréquente par les membres de la communauté et la séparation entre l'analyse (analyse des données) et l'interprétation (ce que signifient les résultats) aident à soutenir la voix des participant·e·s et à accroître la précision et la pertinence de l'analyse.

Les voix des participant·e·s jouent également un rôle potentiel dans l'analyse statistique formelle. Andersson et ses collègues utilisent les poids générés par la cartographie cognitive floue comme *a priori* bayésiens, intégrant les croyances et connaissances préexistantes comme une distribution antérieure des probabilités. Cela permet l'intégration de la perspective autochtone dans l'analyse statistique.

Communication et activités de transfert des connaissances

La diffusion et le transfert des connaissances sont des phases essentielles en évaluation. À ces étapes, il existe un risque élevé d'exploiter les participant·e·s et d'utiliser les résultats pour projeter une situation que les communautés ne pourraient supporter. Au lieu de séparer l'activité de transferts des connaissances et de la considérer comme le produit final de l'évaluation, une évaluation culturellement sûre l'intègre dans le processus d'évaluation (Kothari, McCutcheon, et Graham 2017). Cette approche fait appel à toutes les parties prenantes – le personnel chargé

de l'exécution du programme, les bailleurs de fonds, les participant·e·s et, si possible, les responsables politiques— dans l'évaluation dès le départ. Toutes participent donc à la conception, à la mise en œuvre et à l'interprétation. Andersson et ses collègues ont conçu un protocole appelé SEPA (*socialising evidenced for participatory action*, ou socialisation des données probantes pour l'action participative) qui permet d'intégrer ces étapes dans la recherche. Le protocole SEPA consiste à faire participer les parties prenantes dans la production des données probantes mais aussi à présenter les données de recherche pour qu'elles puissent participer non seulement à leur interprétation mais aussi au développement de solutions dans des dialogues. Les solutions sont ainsi contextualisées, mises en œuvre et évaluées (Ledogar et al. 2017).

II. En quoi cette approche est-elle utile pour l'évaluation des politiques publiques?

La sécurisation culturelle est la première question que l'on devrait se poser en amorçant l'évaluation d'une politique publique. En plus de répondre à des questions d'évaluation *ex ante* et *ex post*, l'approche par sécurisation culturelle contribue à amplifier les voix des participant·e·s et des bénéficiaires. Elle devrait être une exigence pour toutes les évaluations portant sur les services publics. Par exemple, dans les programmes visant à répondre aux besoins des Autochtones ou des minorités, les faire participer à l'évaluation, entendre comment elles et ils définissent le problème (besoins et pertinence de l'évaluation) et comprendre comment elles et ils perçoivent l'adéquation culturelle des solutions potentielles ne peut que renforcer et rendre plus appropriée la politique (Cram 2016; Cameron et al. 2010).

La sécurisation culturelle aide à éviter les obstacles de la mise en œuvre, en augmentant l'efficacité du programme en incluant les visions du monde, les besoins et les priorités des Autochtones (évaluation de l'efficacité). Elle peut contribuer à réduire les inégalités (évaluation d'impact à court terme). Cela peut autonomiser ces communautés et finalement conduire à leur autodétermination (par exemple, une évaluation d'impact à long terme).

III. Un exemple d'utilisation de cette approche en santé reproductive

Un projet de santé reproductive dans une région d'après-conflit du nord de l'Ouganda a utilisé le cadre de sécurisation culturelle pour améliorer les résultats en matière de santé reproductive. Le projet a reçu un financement d'une organisation canadienne (Belaid et al. 2020; Belaid et al. 2021).

La guerre civile (1986-2006) entre l'Armée de Résistance du Seigneur (nord) et le gouvernement de Museveni (sud) a déplacé plus de 90 % de la population de cette région. Cela a accru les tensions de longue date entre le nord et le sud de l'Ouganda (Laruni 2015).

Les communautés du Nord se remettent encore du conflit. La région présente de très mauvais indicateurs de pauvreté, d'opportunités sociales et de développement humain (Esuruku 2019). Le conflit a eu des répercussions négatives sur les services de santé, détériorant la santé maternelle (Chi et al. 2015b, 2015a). Les femmes et les filles de cette région sont moins instruites et plus pauvres. Elles sont beaucoup moins susceptibles de donner naissance dans un établissement de santé (Uganda Bureau of Statistics (UBOS) 2012).

Avant de lancer le projet, nous avons investi du temps dans l'établissement de relations avec les parties prenantes locales et dans le développement de réseaux pour impliquer les membres de la communauté. Les parties prenantes comprenaient des femmes et des hommes de différents groupes d'âge, des sages-femmes traditionnelles, des prestataires de services et des responsables gouvernementaux. Nous avons impliqué ces groupes dans toutes les activités de la conception du programme. Chaque groupe a défini les résultats de soins périnataux en fonction de sa vision du monde. Nous avons utilisé la cartographie cognitive floue pour collecter et comparer les perspectives. Les discussions de groupe ont permis de clarifier le lexique et les concepts culturels associés aux soins périnataux. Nous avons utilisé ces concepts pour concevoir le questionnaire, dans la mesure du possible, en utilisant aussi des questions standardisées correspondant aux concepts identifiés par les parties prenantes.

Les groupes se sont rencontrés dans une série de dialogues délibératifs pour discuter des données probantes locales, générer des listes de stratégies potentielles pour améliorer l'accès aux soins périnataux et concevoir un programme. Nous avons invité les groupes à discuter de qui devrait proposer le programme, selon quelles modalités et quel contenu. Les participant-e-s ont identifié plusieurs obstacles à l'accès aux soins périnataux et ont proposé des stratégies pour résoudre les problèmes d'une manière culturellement sûre.

La sécurisation culturelle a permis d'identifier les problèmes de prestation de services périnataux. La réflexion sur les données probantes locales a généré des solutions faisables menées par la communauté. Par effet de ricochet, cela a accru la confiance entre les membres de la communauté et les prestataires de services.

IV. Quels sont les critères permettant de juger de la qualité de la mobilisation de cette approche?

La sécurisation culturelle ne peut être jugée que par les bénéficiaires du programme (Wilson et Neville 2009; Cameron et al. 2010; CIET/PRAM 2022). Cependant, les équipes d'évaluation peuvent réfléchir aux questions suivantes :

- Les participant·e·s/bénéficiaires déclarent-ils/elles se sentir culturellement en sécurité lors de l'évaluation?
- Comment les bénéficiaires visé·e·s participeront-ils/elles à chaque phase de l'évaluation?
- Les termes de l'évaluation se prêtent-ils à un partenariat?
- L'évaluation s'appuie-t-elle sur les points forts/positifs des communautés?
- L'évaluation augmente-t-elle l'appropriation du projet ou du service, et des produits d'évaluation?
- Comment les méthodes sont-elles adaptées à la culture spécifique?
- Quel est l'impact anticipé de l'évaluation sur l'autodétermination des communautés?

V. Quels sont les atouts et les limites de cette approche par rapport à d'autres?

Un cadre de sécurisation culturelle présente plusieurs avantages pour les équipes d'évaluation et les participant·e·s aux programmes. Elle augmente l'acceptabilité locale et la pertinence de l'évaluation. Elle peut guider la conception des programmes et des services, en augmentant leur

adéquation au contexte. Dans l'évaluation des programmes en cours, la sécurisation culturelle est un prisme interprétatif pour comprendre comment les communautés autochtones et minoritaires vivent ces programmes et services. En tant que prisme analytique, elle peut mettre en évidence la façon dont les inégalités et les injustices sociales sont façonnées, les changements nécessaires et les obstacles ou les facilitateurs possibles de ces changements (Gerlach 2012).

Le principal défi est de développer et de convenir de protocoles pour évaluer l'impact de la sécurisation culturelle dans le contexte de résultats complexes (Gerlach 2012; Tremblay et al. 2020). La sécurisation culturelle dépend dans une large mesure de chaque contexte local, car chaque groupe culturel est différent, a sa propre façon de voir les choses et son propre degré d'adaptation aux représentations dominantes (Cameron et al. 2010). Cependant, à mesure que de plus en plus d'évaluations appliquent un cadre de sécurisation culturelle, nos expériences des meilleures pratiques s'accumuleront et contribueront au développement de lignes directrices avec un large éventail de transférabilité.

Quelques références bibliographiques pour aller plus loin

Andersson, Neil. et Shea, Beverley. et Amaratunga, Carol. et McGuire, Patricia. et Sioui, Georges. 2010. « Rebuilding from Resilience: Research Framework for a Randomized Controlled Trial of Community-led Interventions to Prevent Domestic Violence in Aboriginal Communities. » *Pimatisiwin* 8(2): 61-88.

Andersson, Neil. et Silver, Hilah. 2019. « Fuzzy cognitive mapping: An old tool with new uses in nursing research. » *Journal of Advanced Nursing* 75, no. 12: 3823-30.

- Belaid, Loubna. et Atim, Pamela. et Atim, Eunice. et Ochola, Emmanuel. et Bayo, Pontius. et Oola, Janet. et Sarmiento, Ivan. et Zarowsky, Christina. et Andersson, Neil. 2020. « Marginalized women and services providers improve access to perinatal care in post-conflict Northern Uganda: socializing evidence for participatory action. » *Family Medicine and Community Health* 9:e000610.
- Cameron, Mary. et Andersson, Neil. et McDowell, Ian. et Ledogar, Robert. 2010. « Culturally Safe Epidemiology: Oxymoron or Scientific Imperative. » *Pimatisiwin* 8(2): 89-116.
- Curtis, Elana. et Jones, Rhys. et Tipene-Leach, David. et Walker, Curtis. et Loring, Belinda. et Paine, Sarah-Jane. et Reid, Papaarangi. 2019. « Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition. » *International Journal for Equity in Health* 18(1): 174. <https://doi.org/10.1186/s12939-019-1082-3>. <https://doi.org/10.1186/s12939-019-1082-3>.
- Dion, Anna. et Carini-Gutierrez, Alessandro. et Jimenez, Vania. et Ben Ameur, Amal. et Robert, Emilie. et Joseph, Lawrence. et Andersson, Neil. 2022. « Weight of Evidence: Participatory Methods and Bayesian Updating to Contextualize Evidence Synthesis in Stakeholders' Knowledge. » *J Mix Methods Res* 16(3): 281-306. <https://doi.org/10.1177/15586898211037412>.
- Gerlach, Alison J.. 2012. « A Critical Reflection on the Concept of Cultural Safety. » *Canadian Journal of Occupational Therapy* 79(3): 151-58. <https://doi.org/10.2182/cjot.2012.79.3.4>. <https://journals.sagepub.com/doi/abs/10.2182/cjot.2012.79.3.4>.
- Giles, Brian. et Findlay, C. Scott. et Haas, George. et LaFrance, Brenda. et Laughing, Wesley. et Pembleton, Sakakohe. 2007. « Integrating conventional science and aboriginal perspectives on diabetes using fuzzy cognitive maps. » *Social Science & Medicine* 64(3): 562-76.

<https://doi.org/https://doi.org/10.1016/j.socscimed.2006.09.007>.
<http://www.sciencedirect.com/science/article/pii/S0277953606004758>.

Gollan, Sharon. et Stacey, Kathleen. 2021. *Australian Evaluation Society First Nations Cultural Safety Framework*. Australian Evaluation Society (Melbourne).

Gomez Cardona, Liliana. et Brown, Kristyn. et McComber, Mary. et Outerbridge, Joy. et Parent-Racine, Echo. et Phillips, Allyson. et Boyer, Cyndy. et Martin, Codey. et Splicer, Brooke. et Thompson, Darrell. et Yang, Michelle. et Velupillai, Gajanan. et Laliberté, Arlène. et Haswell, Melissa. et Linnaranta, Outi. 2021. « Depression or resilience? A participatory study to identify an appropriate assessment tool with Kanien'kéha (Mohawk) and Inuit in Quebec. » *Social Psychiatry and Psychiatric Epidemiology* 56(10): 1891-902. <https://doi.org/10.1007/s00127-021-02057-1>. <https://doi.org/10.1007/s00127-021-02057-1>.

Ledogar, Robert. et Arosteguí, Jorge. et Hernández-Alvarez, Carlos. et Morales-Perez, Arcadio. et Nava-Aguilera, Elizabeth. et Legorreta-Soberanis, José. et Suazo-Laguna, Harold. et Belli, Alejandro. et Laucirica, Jorge. et Coloma, Josefina. et Harris, Eva. et Andersson, Neil. 2017. « Mobilising communities for *Aedes aegypti* control: the SEPA approach. » *BMC Public Health* 17(1): 403. <https://doi.org/10.1186/s12889-017-4298-4>. <https://doi.org/10.1186/s12889-017-4298-4>.

Nations, Le Centre de Gouvernance de l'Information des Premières. 1998. « Les principes de PCAP® des Premières Nations. » Le Centre de Gouvernance de l'Information des Premières Nations.

Papps, Elaine. et Ramsden, Irihapeti. 1996. « Cultural Safety in Nursing: the New Zealand Experience. » *International journal for quality in health care: journal of the International Society for Quality in Health Care / ISQua* 8: 491-7. <https://doi.org/10.1093/intqhc/8.5.491>.

- Sarmiento, Ivan. et Paredes-Solís, Sergio. et Loutfi, David. et Dion, Anna. et Cockcroft, Anne. et Andersson, Neil. 2020. « Fuzzy cognitive mapping and soft models of indigenous knowledge on maternal health in Guerrero, Mexico. » *BMC Medical Research Methodology* 20(1): 125. <https://doi.org/10.1186/s12874-020-00998-w>. <https://doi.org/10.1186/s12874-020-00998-w>.
- Smith, Linda Tuhiwai. 2012. *Decolonising methodologies. Research and Indigenous Peoples*, New York: Bloomsbury.
- Smye, Vicky. et Browne, Annette. 2002. « 'Cultural safety' and the analysis of health policy affecting aboriginal people. » *Nurse Res* 9(3): 42-56. <https://doi.org/10.7748/nr2002.04.9.3.42.c6188>.
- Williams, Robyn. 1999. « Cultural safety — what does it mean for our work practice? » *Australian and New Zealand Journal of Public Health* 23(2): 213-14. <https://doi.org/https://doi.org/10.1111/j.1467-842X.1999.tb01240.x>. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1999.tb01240.x>.
- Wilson, Denise. et Neville, Stephen. 2009. « Culturally safe research with vulnerable populations. » *Contemporary Nurse* 33(1): 69-79. <https://doi.org/10.5172/conu.33.1.69>. <https://doi.org/10.5172/conu.33.1.69>.

Présentation des auteur·rice·s

Mathias André est diplômé de l'École nationale de la statistique et de l'administration économique (Ensaie), docteur en économie de l'École Polytechnique et administrateur de l'Insee. Expert est en charge de l'élaboration des comptes nationaux distribués au sein de l'Insee, il a été rapporteur du groupe d'experts sur la mesure de la redistribution et des inégalités publié en février 2021 et auteur d'un rapport sur les enjeux redistributifs pour le Conseil des prélèvements obligatoires. Il a publié des travaux sur la microsimulation des réformes socio-fiscales et le rôle des différents transferts dans la réduction des inégalités (TVA, impôt sur le revenu, taxe foncière). Il mène également des études sur le patrimoine immobilier ou les inégalités et le climat. Il assure un enseignement sur la microsimulation à l'Ensaie.

mathias.andre@insee.fr

Neil Andersson MD PhD est professeur de médecine familiale et directeur de la recherche participative à l'Université McGill à Montréal, Canada. Il se spécialise dans les essais communautaires à grande échelle en mettant l'accent sur l'engagement des patient·e·s et de la communauté. Les interventions étudiées ont inclus la prévention des maladies à transmission vectorielle, la vaccination, la sécurisation culturelle de l'accouchement, les visites périnatales à domicile, la prévention de la violence sexiste et l'accès au soutien social pour les jeunes femmes à risque de VIH. Ses contributions à la méthodologie des essais ont mis l'accent sur l'engagement précoce des parties prenantes dans la conceptualisation et la conception des interventions, et l'analyse par grappes des résultats. Il plaide pour une inclusion éthique et culturellement sûre dans les essais contrôlés randomisés de groupes socialement marginalisés et autochtones.

neil.andersson@mcgill.ca

Habibata Baldé est une chercheuse en sciences sociales avec une spécialisation en Sciences Sociales Appliquées à la santé et une expertise en recherche qualitative. Depuis 2014, elle est activement impliquée dans des activités de recherche en santé. Elle est actuellement Assistante Technique au Centre d'Excellence d'Afrique pour la Prévention et le Contrôle des Maladies Transmissibles de l'Université de Conakry et consultante de la Guinée dans un projet de recherche multi pays sur l'introduction de l'Oxymètre de Pouls dans les centres de santé primaires avec l'ONG ALIMA.

habicompanya@gmail.com

Carlo Barone est professeur titulaire de sociologie au CRIS, Sciences Po, Paris. Il est le directeur de l'axe de recherche « politiques éducatives » du LIEPP, et il est affilié au J-PAL. Il mène plusieurs évaluations d'impact des politiques dans le domaine de l'éducation en utilisant des essais contrôlés randomisés ainsi que des revues systématiques et des méta-analyses. Il a publié plusieurs articles sur les inégalités éducatives, les inégalités sur le marché du travail et la mobilité sociale dans des revues de sociologie de premier plan.

carlo.barone@sciencespo.fr

Loubna Belaid est professeure adjointe en évaluation de programmes à l'École Nationale d'Administration Publique de Montréal, Canada. Ses intérêts de recherche portent sur la sécurisation culturelle et la justice sociale dans les programmes et politiques de santé. Elle utilise des approches participatives pour co-concevoir, mettre en œuvre et évaluer des programmes de santé avec et pour les communautés autochtones, racialisées et marginalisées au Canada et en Afrique de l'Est.

Loubna.Belaid@enap.ca

Genowefa Blundo-Canto est chercheuse en sciences sociales au Centre de recherche agronomique pour le développement international (Cirad). Auparavant, elle a travaillé comme consultante et chargée de recherche

à Bioversity International et comme post-doc en évaluation d'impact au Centre international d'agriculture tropicale (CIAT). Économiste du développement, elle mène des recherches sur l'évaluation d'impact des interventions de recherche agricole pour le développement (AR4D). Elle s'intéresse particulièrement aux questions méthodologiques de l'évaluation d'impact, en se concentrant sur les méthodes mixtes intégrées, les approches participatives et systémiques, et la navigation dans la complexité à des échelles multiples. Elle combine également la prospective participative avec des méthodes d'évaluation d'impact qualitatives et quantitatives afin d'élargir la portée des évaluations. Sur le plan thématique, elle se concentre sur les impacts multidimensionnels et multi-échelles des interventions qui améliorent l'utilisation de la biodiversité agricole, avec un intérêt particulier pour la sécurité alimentaire et nutritionnelle, l'équité sociale et les déséquilibres de pouvoir à plusieurs échelles.

genowefa.blundo_canto@cirad.fr

Abdourahmane Coulibaly est anthropologue, enseignant-chercheur à la Faculté de médecine et d'odontostomatologie (Mali) et membre de l'IRL « Environnement – Santé – Sociétés » UCAD, USTTB, CNRST – Ouagadougou, CNRS – France. Il a participé à plusieurs programmes de recherche portant sur la mise en œuvre des politiques de santé notamment le financement basé sur les résultats, les innovations technologiques dans le domaine de la santé ou encore l'analyse de la résilience des institutions sanitaires. Ses recherches sont largement basées sur l'utilisation des méthodes qualitatives et la démarche ethnographique.

coulibalyabdourahmane@gmail.com

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*.

tdelahais@quadrant-conseil.fr

Agathe Devaux-Spatarakis est consultante et chercheuse pour la Scop Quadrant Conseil. Elle conduit des missions d'évaluation de politiques publiques et d'accompagnement méthodologique pour le compte d'organisations publiques ou d'ONG en France et à l'international. Docteure en science politique, elle est spécialisée dans le développement des méthodes d'évaluation adaptées aux innovations et expérimentations sociales ainsi que l'étude de l'utilisation des résultats d'évaluation par les responsables politiques.

adevaux@quadrant-conseil.fr

Emanuele Ferragina est professeur associé de sociologie à Sciences Po, Paris. Avant sa nomination à Paris, il a travaillé comme chercheur et maître de conférences à l'université d'Oxford. Emanuele a étudié à Turin, Bordeaux, Londres et Paris et a obtenu son doctorat en politique sociale à l'université d'Oxford. Ses travaux ont été publiés dans plusieurs revues d'économie politique, de sociologie, de sciences politiques et de politique sociale, telles que la *Review of International Political Economy*, *New Political Economy*, *Research in Social Stratification & Mobility*, *Socius*, *International Journal of Comparative Sociology*, *Political Studies Review*, *Journal of European Social Policy*, *Social Policy & Administration*, *Social Politics*, *Social Policy & Society*, *Stato & Mercato*, *L'Année Sociologique*.

emanuele.ferragina@sciencespo.fr

Nicolas Fischer est chargé de recherche CNRS en science politique au Centre de recherches sociologiques sur le droit et les institutions pénales (CESDIP). Ses recherches récentes ont porté sur la détention

administrative des étrangers en France et les politiques d'immigration, sur le contrôle indépendant des lieux d'enfermement, et plus largement sur la tension entre répression violente et protection juridique des populations stigmatisées dans les démocraties. Il mène actuellement une enquête sur les controverses médico-judiciaires visant les exécutions capitales par injection létale aux États-Unis. Il a notamment publié *Le territoire de l'expulsion. La rétention administrative des étrangers et l'Etat de droit en France* (Lyon: ENS Éditions, 2017) et, avec Camille Hamidi, *Les politiques migratoires* (Paris: la Découverte coll. Repères, 2016).

fischer@cesdip.fr

Denis Fougère est Directeur de Recherches de classe exceptionnelle au CNRS. Il est membre du Centre de Recherches sur les Inégalités Sociales (CRIS) et du Laboratoire Interdisciplinaire d'Évaluation des Politiques Publiques (LIEPP) à Sciences Po Paris. Il enseigne l'économie de l'éducation et les méthodes statistiques d'évaluation des politiques publiques à Sciences Po Paris. Il est également Research Fellow du Centre for Economic Policy Research (CEPR, Londres) et de l'Institute of Labor Economics (IZA, Bonn). Ses recherches actuelles portent sur l'évaluation des politiques éducatives et des réformes des systèmes de retraite en France. Il a publié des articles dans plusieurs revues internationales, telles qu'*Econometrica*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Economic Journal*, *European Economic Review*, *European Sociological Review*, *Journal of Public Economics*, *Journal of the European Economic Association*, *Journal of Business and Economic Statistics*, *Journal of Applied Econometrics*, *The Econometrics Journal*, *Journal of Population Economics*, *Labour Economics*, etc.

denis.fougere@sciencespo.fr

Lara Gautier est professeure adjointe à l'École de santé publique de l'Université de Montréal et chercheuse régulière au Centre de recherche en santé publique, et à l'Institut de recherche SHERPA, à Montréal

(Canada). Formée en santé publique et en sciences politiques, elle possède une expertise scientifique en évaluation participative des services de santé employant les méthodes de recherche qualitatives et mixtes.

lara.gautier@umontreal.ca

Pauline Givord est depuis décembre 2022 cheffe du département des études économiques de l'Institut national de la statistique et des études économiques (Insee). Auparavant, elle a occupé plusieurs postes au sein de l'Insee, centrés sur les études économiques sur des sujets très variés et la méthodologie statistique. Experte en méthodes économétriques d'évaluation des politiques publiques, elle a également participé à la création au sein de l'Insee du SSP-Lab, visant à promouvoir l'innovation en matière de sources de données et de méthodes de data science, relatives aux productions statistiques du système de statistique publique (SSP). Elle est également passée par le Centre de recherche en économie et en statistique (Crest), l'OCDE, et la direction de l'animation de la recherche (Dares) au sein du Ministère du travail, où elle était en charge du suivi de l'évaluation du Plan d'Investissement dans les compétences. Elle est diplômée de l'Ecole Polytechnique et de l'ENSAE, titulaire d'un doctorat en sciences économiques et d'une habilitation à diriger des recherches.

pauline.givord@insee.fr

Charlotte Halpern est titulaire d'un doctorat en sciences politiques, chercheuse titulaire en science politique au Centre d'études européennes et de politique comparée (CEE) de Sciences Po et codirectrice du groupe de recherche sur les politiques environnementales au Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP). Ses travaux examinent les processus de changement dans les politiques publiques et la relation entre les mobilisations sociales et les dynamiques de restructuration de l'État. Ses recherches actuelles portent sur la gouvernance des politiques de transition durable dans les villes européennes. Ses publications récentes comprennent des numéros spéciaux et des articles dans des revues (par exemple, *Comparative*

European Politics; West European Politics; Politique européenne...) et deux volumes édités, *Policy analysis in France* (Policy press, 2018 co-éd. avec P. Hassenteufel et P. Zittoun) et *Villes sobres* (Presses de Sciences Po, 2018). Elle est directrice scientifique de l'Executive master de Sciences Po sur la gouvernance territoriale et l'urbanisme, et enseigne l'analyse comparée des politiques publiques, la gouvernance urbaine et les politiques environnementales à Sciences Po et AgroParisTech.

charlotte.halpern@sciencespo.fr

Quan Nha Hong est professeure adjointe à l'École de réadaptation de l'Université de Montréal et chercheuse au Centre de recherche interdisciplinaire en réadaptation du Montréal métropolitain (CRIR) – Institut universitaire sur la réadaptation en déficience physique de Montréal (IURDPM). Elle est ergothérapeute avec des formations de recherche en sciences cliniques (M.Sc., Université de Sherbrooke), en évaluation des technologies de la santé (M.Sc., Université de Montréal), et en soins de première ligne (Ph.D., Université McGill). Elle a aussi complété un stage postdoctoral au Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) de l'University College London (UCL) et à l'Institut national d'excellence en santé et en services sociaux (INESSS). Elle s'intéresse particulièrement aux méthodes de recherche, dont les revues systématiques et les méthodes mixtes, ainsi qu'au transfert de connaissances pour favoriser la prise de décision informée par les données probantes.

quan.nha.hong@umontreal.ca

Nicolas Jacquemet est Professeur d'économie à l'Université Paris 1 Panthéon-Sorbonne et à PSE-Ecole d'Economie de Paris, où il dirige le master économie et psychologie. Il a collaboré à la rédaction de différents manuels à destination d'étudiants de master ou de professionnels de l'évaluation portant sur l'économétrie (De Boeck) et sur les méthodes expérimentales de l'économie (Cambridge University Press, *Economica*). Ses travaux en matière d'évaluation des politiques publiques combinent

diverses méthodes expérimentales (testing, expériences de terrain, expériences contrôlées en laboratoire) ainsi que des données d'enquête ou administratives et portent notamment sur la discrimination à l'embauche, l'évasion fiscale, le pilotage de l'offre de soins ou encore le partage des tâches à l'intérieur des couples. Il contribue régulièrement au débat public, et a collaboré notamment à la rédaction d'un ouvrage grand public présentant les contributions de l'économie comportementale à la réflexion en matière d'élaboration des politiques publiques (La découverte, collection Repères).

nicolas.jacquemet@univ-paris1.fr

Sarah Louart est doctorante en socio-économie de la santé, rattachée au Centre lillois d'études et de recherches sociologiques et économiques (Clersé, Université de Lille). Elle étudie notamment les questions de l'accès aux soins pour les populations démunies et les processus d'introduction et de diffusion des innovations en santé. Elle travaille actuellement avec l'ONG ALIMA et l'Institut de Recherche pour le Développement (IRD) à Dakar, sur l'évaluation réaliste d'un projet visant à diffuser une innovation en santé dans les centres de santé primaire de quatre pays d'Afrique de l'Ouest.

sarah.louart@gmail.com

Ana Manzano est professeure associée en politique publique à l'Université de Leeds (Royaume-Uni) et experte en sciences sociales appliquées dans le domaine de l'évaluation et de la politique, des systèmes et des pratiques de soins de santé. Ses domaines d'expertise sont l'évaluation réaliste, les méthodes qualitatives avancées et la relation entre méthodes, données probantes et évaluation de programmes. Elle a travaillé sur des études d'évaluation internationales en Europe, en Afrique et en Asie du Sud. Manzano publie régulièrement dans les principales revues spécialisées dans l'évaluation, la méthodologie et les politiques et systèmes de santé, notamment *Evaluation*, *Implementation Science*, *Evaluation and Program Planning*, *the International Journal of Social*

Research Methodology, Social Sciences and Medicine, et Health Policy and Planning. Elle a participé au projet RAMESES II, qui a élaboré des normes pour la réalisation d'évaluations réalistes. Elle est également directrice des méthodes qualitatives avancées au White Rose Social Sciences Doctoral Training Partnership (Royaume-Uni).

A.Manzano@leeds.ac.uk

Valérie Pattyn est professeure associée à l'Institut d'administration publique de l'Université de Leiden (Pays-Bas), et est partiellement affiliée à l'Institut de gouvernance publique de la KU Leuven (Belgique). Ses travaux portent sur les politiques publiques, l'élaboration de politiques informées par des données probantes et l'évaluation. Son programme de recherche actuel porte sur des questions telles que le développement de systèmes d'évaluation et l'impact sur les décisions politiques ; la production et l'utilisation de données probantes et de conseils politiques au sein et en dehors de la fonction publique ; et l'élaboration des politiques publiques dans des conditions d'incertitude. Outre la recherche fondamentale, elle a acquis une expérience considérable en matière de conseil politique et de recherche en évaluation dans divers domaines. Elle est membre de plusieurs réseaux nationaux et internationaux, dont le Groupe européen d'administration publique (coprésidente du groupe d'étude permanent sur la conception et l'évaluation des politiques), l'Association flamande d'évaluation (comité de coordination), l'Association néerlandaise d'évaluation (groupe central) et COMPARative Methods for Systematic cross-caSe analysis (comité consultatif).

v.e.pattyn@fgga.leidenuniv.nl

Clément Pin est maître de conférences en sociologie à l'INSEI, membre du GRHAPES, chercheur affilié au LIEPP (Sciences Po) et au laboratoire EMA (Cergy Paris Université). Ses travaux portent sur : 1) les politiques éducatives et universitaires, 2) les dispositifs et instruments d'orientation scolaire et professionnelle, 3) les médiations locales (relations écoles-

familles et plus largement gouvernance territoriale), 4) les méthodes d'évaluation (qualitatives et mixtes). Il a récemment publié avec Carlo Barone « L'apport des méthodes mixtes à l'évaluation » (*Revue française de science politique*, 2021/3) et avec Agnès van Zanten « The Impact on French Upper Secondary Schools of Reforms Aiming to Improve Students' Transition to Higher Education » (*Oxford Research Encyclopedia of Education*, OUP, 2021).

clement.pin@sciencespo.fr

Pierre Pluye (†) était professeur au Département de médecine familiale de l'Université McGill, et membre associé de l'École des sciences de l'information. Il a codirigé l'association « Méthodes mixtes francophonie » (MMF) et l'axe « Valorisation des données » de l'Unité Système de Santé Apprenant Québec. Il était membre de l'Académie canadienne des sciences de la santé, et membre fondateur entre autres du Réseau québécois de recherche axé sur les pratiques de première ligne. En 2017, il a reçu le prix « Chercheur de l'année » du Collège des médecins de famille du Canada. En 2021, il a reçu le « Doctoral Teaching Award » de l'association des universités du nord-est (Canada/États-Unis) qui reconnaît l'excellence et l'innovation dans l'enseignement doctoral. Par son expertise en méthodes mixtes et revues de littérature mixtes, il a considérablement contribué au développement de ces méthodes. Ses derniers travaux de recherche visaient à évaluer et à améliorer les effets de l'information en ligne sur la santé. La carrière du Professeur Pierre Pluye, décédé le 1er août 2023, est retracée ici par le Comité directeur du collectif Méthodes mixtes francophonie : http://methodesmixtesfrancophonie.pbworks.com/w/file/fetch/154045617/PP_2023-08-01.pdf

Estelle Raimondo est Conseillère en méthodes au Groupe indépendant d'évaluation (Independent evaluation group, ou IEG) de la Banque mondiale, où elle dirige des évaluations et conseille les équipes de l'IEG sur la conception et les innovations méthodologiques. Forte de plus de dix ans d'expérience en évaluation des politiques de développement, elle

est membre de l'équipe professorale du Programme international de formation à l'évaluation du développement (International Program for Development Evaluation Training, ou IPDET) et siège au conseil d'administration de la Société européenne d'évaluation. Ses recherches ont été publiées dans plusieurs revues et ouvrages internationaux évalués par des pairs. Elle a obtenu son doctorat en évaluation à l'Université George Washington et un double master en politique économique internationale de Sciences Po Paris et de l'Université Columbia.

eraimondo@worldbank.org

Thomas Rapp, Maître de conférences (HDR) en sciences économiques au LIRAES (Université Paris Cité), codirecteur de l'axe « Politiques de santé » du LIEPP, et titulaire de la Chaire AgingUP! Il est spécialisé en économie de la santé, économie du vieillissement et en analyse des politiques de santé. Il est l'auteur de plus de 70 publications (articles, rapports, chapitres d'ouvrage) sur ces thèmes. Il a été Harkness fellow à Harvard (2015-2016), économiste de la santé à l'OCDE (2017-2019), et professeur invité à Harvard, Columbia et à l'Université Catholique de Rome. Il est éditeur associé de la revue scientifique *Value in Health*. Depuis 10 ans, son programme de recherches a reçu l'appui de plusieurs financements, notamment obtenus auprès l'Agence nationale de la recherche, du Commonwealth Fund de New York, et du programme Innovative Medicines Initiative de la Commission Européenne.

thomas.rapp@u-paris.fr

Anne Revillard est professeure associée en sociologie à Sciences Po, membre du Centre de recherche sur les inégalités sociales (CRIS) et directrice du Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP). Ses recherches portent sur l'articulation entre droit, action publique et transformations contemporaines des systèmes d'inégalités liées au genre et au handicap. Elle contribue notamment aux réflexions en évaluation à partir d'une focale sur les méthodes qualitatives

et d'une approche en termes de réception de l'action publique, délibérément centrée sur le point de vue des ressortissant-e-s de la politique étudiée.

anne.revillard@sciencespo.fr

Valéry Ridde est directeur de recherche au CEPED, une unité de recherche commune à l'Université de Paris et à l'Institut de recherche pour le développement (IRD). Il est actuellement basé à l'Institut de la santé et du développement (ISED) de l'Université Cheikh Anta Diop de Dakar (Sénégal). Ses travaux de recherche et d'évaluation portent sur la couverture maladie universelle, le financement des services de santé, l'évaluation des programmes, les politiques de santé publique et la promotion de la santé.

valery.ridde@ird.fr

Émilie Robert est professeure associée à l'École de santé publique de l'Université de Montréal. Spécialiste de l'approche réaliste pour l'évaluation et la synthèse des connaissances, elle forme les équipes de recherche et les accompagne pour la conception et la mise en œuvre de leurs projets de recherche évaluative. Émilie a réalisé des mandats pour plusieurs organisations internationales et provinciales, publiques et non-gouvernementales, et universitaires. Son approche est ancrée dans l'évaluation fondée sur l'utilisation des connaissances et le développement d'une pensée évaluative.

emilierobert.udem@gmail.com

Lou Safra est titulaire d'une thèse en sciences cognitives de l'École normale supérieure (Paris) et enseignante-chercheuse à Sciences Po depuis 2018. Membre du CEVIPOF, elle est également associée à l'Institut d'Études Cognitives (Laboratoire de Neurosciences Cognitives & Laboratoire de Neurosciences Cognitives et Computationnelles, École Normale Supérieure, Paris) et affiliée au Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP). Dans ses recherches, elle

applique les concepts et les méthodes de sciences cognitives à l'étude des comportements politiques et sociaux. Elle s'intéresse plus particulièrement aux causes et aux conséquences des inégalités sociales, politiques et économiques, en analysant à la fois les réactions face à ces inégalités, les effets comportementaux de ces inégalités sur les populations défavorisées et la façon dont les politiques publiques peuvent contribuer à accentuer ou à limiter ces inégalités. Pour cela, elle combine l'utilisation d'expérimentations en laboratoire à l'analyse de données d'enquête internationales et d'objets culturels.

lou.safra@sciencespo.fr

Remerciements

Cet ouvrage est le résultat de la dynamique scientifique impulsée depuis 2011 par le Laboratoire interdisciplinaire d'évaluation des politiques publiques (LIEPP), où s'est créé au fil des recherches un espace de dialogue entre différentes traditions méthodologiques et approches évaluatives. Fortement porté par la communauté de chercheuses et de chercheurs impliqué-e-s dans les activités du LIEPP, ce livre vise à faciliter et nourrir cet échange. Je tiens donc en premier lieu à remercier les 25 autrices et auteurs qui, déjà engagé-e-s dans le projet du LIEPP ou sollicité-e-s pour l'occasion, ont donné de leur temps pour préparer les chapitres didactiques qui composent cet ouvrage, en se pliant à une trame de questionnement commune et à de fortes exigences sur le plan du format rédactionnel, donnant lieu à de nombreuses réécritures. Amorcé à l'été 2022, ce projet de publication a été mené à bien en moins d'un an grâce à l'engagement enthousiaste et à la réactivité de contributrices et contributeurs attaché-e-s à communiquer leur passion pour la recherche dans des termes accessibles, dans une démarche de médiation scientifique et de dialogue inter-méthodes.

Le LIEPP a également apporté son soutien financier à ce projet éditorial, permettant la publication en accès ouvert. À ce titre, cet ouvrage a 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 Paris Cité (ANR-18-IDEX-0001).

Au LIEPP, les textes ont été traduits et ont fait l'objet d'une première mise en forme par Konstantinos Papadopoulos, qui a été une cheville ouvrière de ce projet et que je remercie pour son implication. Ariane Lacaze a ensuite fourni un important travail éditorial et de suivi des relectures d'épreuves par les autrices et auteurs : un grand merci à elle

pour son professionnalisme et son efficacité. Merci également à Samira Jebli, Andreana Khristova et Latifa Lousao, de l'équipe du LIEPP, qui ont apporté plus ponctuellement leur concours à ce projet.

Je remercie sincèrement les Éditions science et bien commun, et tout particulièrement Érika Nimis, pour avoir accueilli cette double publication (en français et en anglais), pour leur réactivité et le soin apporté à la préparation de l'ouvrage.

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](https://twitter.com/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

MÉTHODES ET APPROCHES EN ÉVALUATION DES POLITIQUES PUBLIQUES

sous la direction d'Anne Revillard

En tant que pratique de recherche appliquée, l'évaluation des politiques publiques a emprunté toute une série de méthodes aux sciences sociales. Mais son essor a aussi suscité le développement d'approches spécifiques.

Partant de ce constat, deux choix fondamentaux guident cet ouvrage : combiner des outils issus de la recherche fondamentale avec d'autres développés dans la pratique de l'évaluation, et ouvrir un dialogue entre méthodes quantitatives et qualitatives.

24 méthodes ou approches qualitatives, quantitatives ou mixtes font ainsi l'objet de présentations didactiques et illustrées, à partir d'une trame de questionnement commune facilitant leur comparaison.

Par son accessibilité, cet ouvrage constitue aussi bien un outil de dialogue interdisciplinaire et inter-méthodes pour les universitaires, qu'une introduction aux enjeux méthodologiques de l'évaluation pour les étudiant-e-s, praticien-ne-s, les acteurs publics et la société civile.

Anne Revillard est professeure associée en sociologie à Sciences Po (CRIS-LIEPP).



Un accès libre et universel, par le biais du numérique, à des livres scientifiques et documentaires publiés par des auteurs et des autrices de pays des Suds et du Nord, dans la perspective de la justice cognitive et de l'écologie des savoirs. Acheter un livre imprimé, c'est nous soutenir!

editionscienceetbiencommun.org

SciencesPo
LABORATOIRE INTERDISCIPLINAIRE
D'ÉVALUATION DES POLITIQUES PUBLIQUES



ISBN : 978-2-925128-27-4

Design de couverture : Kate McDonnell

Imprimé en France, 2023



Ce livre est aussi disponible
en libre accès sous licence
Creative Commons CC-BY-SA 4.0

