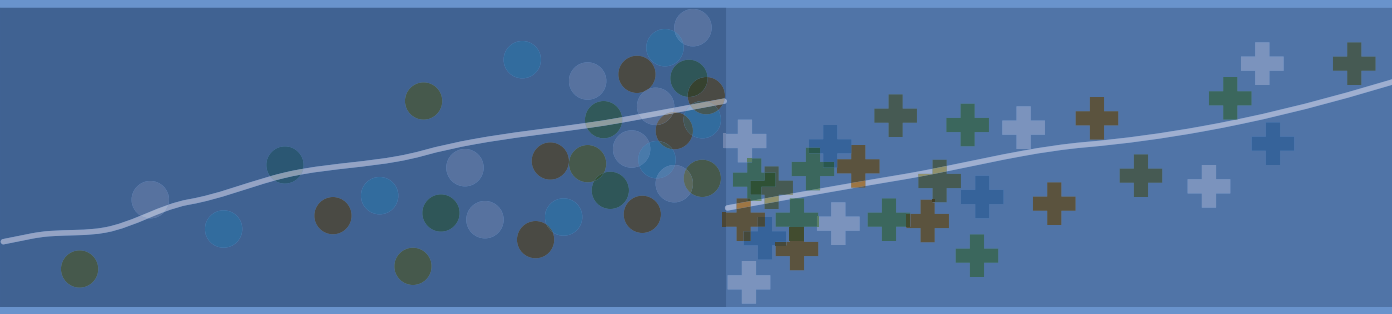
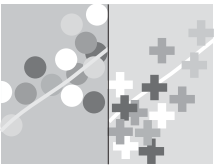
L'évaluation d'impact en pratique



Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
Christel M. J. Vermeersch

BANQUE MONDIALE

L'évaluation d'impact en pratique



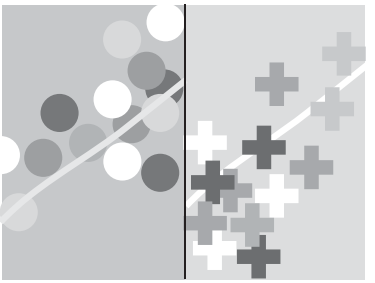
La version anglaise de *l'évaluation d'impact en pratique* est disponible sous la forme d'un manuel interactif à l'adresse suivante : **http://www.worldbank.org/pdt.** La version électronique permet à la communauté des praticiens qui travaillent dans des régions ou des secteurs différents ainsi qu'aux étudiants et aux enseignants de partager des informations et des notes qui favorisent l'apprentissage multimédia et les échanges de connaissances.

Des documents complémentaires au présent ouvrage sont disponibles à l'adresse suivante :
http://www.worldbank.org/ieinpractice.

Ce manuel a été rendu possible par le soutien du Fonds espagnol d'évaluation d'impact (SIEF). Lancé en 2007 avec un don de 14,9 millions de dollars du Gouvernement espagnol et complémenté par un don de 2,1 millions de dollars du département britannique du développement international (DFID), le SIEF est le plus grand fonds consacré à l'évaluation d'impact jamais mis en place par la Banque mondiale. Son objectif principal est de renforcer l'ensemble des preuves existantes sur les programmes qui fonctionnent en matière de santé, d'éducation et de protection sociale et, de ce fait, d'informer les décisions sur les politiques de développement.

http://www.worldbank.org/sief.

L'évaluation d'impact en pratique



Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
Christel M. J. Vermeersch

BANQUE MONDIALE

1818 H Street NW
Washington DC 20433
Téléphone : 202-473-1000
Internet: www.worldbank.org

1 2 3 4 14 13 12 11

Cet ouvrage a été réalisé par le personnel de la Banque internationale pour la reconstruction et le développement/Banque mondiale. Les observations, interprétations et conclusions qu'il contient ne reflètent pas nécessairement l'opinion du Conseil d'administration de la Banque mondiale ou des pays qu'il représente.

La Banque mondiale ne garantit pas l'exactitude des données contenues dans cet ouvrage. Les frontières, les couleurs, les dénominations et toute autre information figurant sur les cartes du présent ouvrage n'impliquent de la part de la Banque mondiale aucun jugement quant au statut juridique d'un territoire quelconque et ne signifient nullement que l'institution reconnaît ou accepte ces frontières.

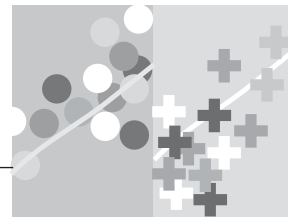
ISBN: 978-0-8213-8752-8
eISBN: 978-0-8213-8681-1
DOI: 10.1596/978-0-8213-8752-8

Données de catalogage avant publication de la Bibliothèque du Congrès

L'évaluation d'impact en pratique/Paul J. Gertler ... [et al.].
p. cm.
Comprend des références bibliographiques et un index.
ISBN 978-0-8213-8541-8 -- ISBN 978-0-8213-8593-7 (électronique)
1. Projets de développement économique--Évaluation. 2. Étude d'évaluation (programmes d'action sociale) I. Gertler, Paul, 1955- II. Banque mondiale.
HD75.9.I47 2010
338.90072--dc22

2010034602

Maquette de couverture : Naylor Design.

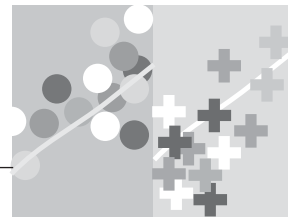


TABLES DES MATIÈRES

Encadrés

Figures

Tableaux



PRÉFACE

Ce manuel constitue une introduction accessible à l'évaluation d'impact et à sa pratique dans le domaine du développement. Il est principalement destiné aux professionnels du développement et aux décideurs, mais peut également être utile aux étudiants et à toute personne intéressée à l'évaluation d'impact. Les évaluations d'impact prospectives visent à déterminer si un programme a atteint ou non les résultats espérés ou à tester différentes stratégies pour atteindre ces résultats. Nous considérons qu'une augmentation du nombre d'évaluations et une amélioration de leur qualité permettront de renforcer l'ensemble des preuves existantes au sujet de l'efficacité des politiques et programmes de développement dans le monde. Notre espoir est que les gouvernements et les professionnels du développement puissent prendre des décisions fondées sur des résultats éprouvés, tels que les preuves générées par les évaluations d'impact, de manière à rendre plus efficace l'utilisation des ressources pour réduire la pauvreté et améliorer le bien-être des populations. Les trois parties du manuel constituent une introduction non technique à l'évaluation d'impact. Elles décrivent ce qu'il convient d'évaluer et pourquoi (partie 1) ; exposent des méthodes d'évaluation (partie 2) ; et indiquent comment mettre en œuvre une évaluation (partie 3). Ces étapes constituent des éléments essentiels à la réalisation d'une évaluation d'impact.

L'approche de l'évaluation d'impact que nous privilégions dans ce manuel est largement intuitive et nous essayons de minimiser les aspects techniques. Nous présentons au lecteur une gamme d'outils d'évaluation d'impact (les concepts et méthodes sous-jacents à toute évaluation d'impact) et illustrons leur application à de réels programmes de développement. Les méthodes évoquées sont directement issues de la recherche appliquée en sciences sociales et ont de nombreux points communs avec les méthodes de recherche utilisées en sciences naturelles. En ce sens, l'évaluation d'impact combine les outils de recherche empiriques couramment utilisés en économie et dans d'autres sciences sociales avec les réalités opérationnelles et politico-économiques de la mise en œuvre de politiques et de pratiques de développement.

D'un point de vue méthodologique, notre approche est essentiellement pragmatique : nous estimons que la méthode d'évaluation la plus pertinente doit être définie

en fonction du contexte opérationnel, et non le contraire. En ce sens, il est essentiel d'intégrer des évaluations d'impact prospectives à la mise en œuvre des projets dès leur conception. Au-delà de la méthode, il est tout aussi important de créer un consensus parmi les parties prenantes à un programme et d'élaborer une évaluation en adéquation avec le contexte politique et opérationnel. Par ailleurs, il nous semble primordial d'être transparent par rapport aux limites des évaluations d'impact. Finalement, nous encourageons vivement les décideurs et les responsables de programme à considérer l'évaluation d'impact à partir d'un cadre logique mettant clairement en évidence les relations causales à travers lesquelles un programme produit des extrants et influence les résultats finaux. Complémenter les évaluations d'impact avec des données de suivi et des évaluations d'autres types permet aussi de mieux appréhender la performance d'un programme.

L'originalité du présent manuel réside surtout dans son approche visant à illustrer l'application des outils d'évaluation d'impact à la réalité des programmes de développement. Nos expériences et observations relatives à la mise en pratique d'évaluations d'impact découlent de notre travail de formation et de collaborations avec des centaines de partenaires chevronnés issus d'institutions publiques, d'universités et d'organisations actives dans le domaine du développement. Entre les auteurs, le manuel tire ainsi parti de dizaines d'années d'expérience dans la réalisation d'évaluation d'impact à travers le monde.

Ce livre est fondé sur une série de ressources pédagogiques mises au point pour les ateliers « Turning Promises to Evidence », organisés par le bureau de l'économiste en chef pour le développement humain, en partenariat avec les unités régionales et le groupe de recherche en économie du développement de la Banque mondiale. Au moment de la rédaction du présent ouvrage, ces ateliers se sont tenus plus d'une vingtaine de fois dans toutes les régions du globe. Tant le manuel que les ateliers ont été réalisés grâce aux généreuses contributions du Gouvernement espagnol et du département britannique du développement international (DfID) par le biais du Fonds espagnol pour l'évaluation d'impact (SIEF). Le manuel, des présentations et des documents complémentaires sont disponibles sur le site http://www.worldbank.org/ieinpractice.

D'autres ressources de qualité proposent une introduction à l'évaluation d'impact, notamment Baker 2000 ; Ravallion 2001, 2008, 2009 ; Duflo, Glennerster et Kremer 2007 ; Duflo et Kremer 2008 ; Khandker, Koolwal et Samad 2009 ; ainsi que Leeuw et Vaessen 2009. La particularité du présent manuel est qu'il combine une revue non technique des méthodes d'évaluation quantitatives tout en établissant un lien direct avec les règles opérationnelles des programmes et en abordant de nombreux aspects pratiques liés à la réalisation d'évaluations. Il est complété par des outils didactiques au sujet de l'évaluation d'impact.

Les ressources pédagogiques sur lesquelles repose le manuel ont été enseignées et améliorées par de nombreux experts renommés ayant tous laissé leur empreinte et leur perspective sur les méthodes d'évaluation d'impact. Paul Gertler, Sebastian Martinez, Sebastian Galiani et Sigrid Vivo ont compilé une première version de ces

ressources pour un atelier organisé par le ministère mexicain du Développement social (SEDESOL) en 2005. Christel Vermeersch a développé et reformulé des sections importantes des modules techniques et adapté une étude de cas pour les besoins de l'atelier. Laura Rawlings et Patrick Premand ont développé des ressources utilisées dans les versions plus récentes de l'atelier.

Nous souhaitons remercier de nombreuses personnes qui ont assuré des formations dans le cadre de l'atelier pour leur importante contribution, en particulier Felipe Barrera, Sergio Bautista-Arredondo, Stefano Bertozzi, Barbara Bruns, Pedro Carneiro, Nancy Qian, Jishnu Das, Damien de Walque, David Evans, Claudio Ferraz, Jed Friedman, Emanuela Galasso, Sebastian Galiani, Gonzalo Hernández Licona, Arianna Legovini, Phillippe Leite, Mattias Lundberg, Karen Macours, Juan Muñoz, Plamen Nikolov, Berk Özler, Gloria M. Rubio et Norbert Schady. Nous remercions également Barbara Bruns, Arianna Legovini, Dan Levy et Emmanuel Skoufias pour leur revue critique d'une version préliminaire de ce manuel, tout comme Bertha Briceno, Gloria M. Rubio et Jennifer Sturdy pour leurs commentaires. Nous tenons également à saluer la grande qualité du travail de l'équipe d'organisation de l'atelier, en particulier Paloma Acevedo, Theresa Adobea Bampoe, Febe Mackey, Silvia Paruzzolo, Tatyana Ringland, Adam Ross, Jennifer Sturdy et Sigrid Vivo.

Ce manuel repose sur des transcriptions de présentations effectuées lors d'un atelier tenu à Beijing en Chine en juillet 2009. Nous remercions toutes les personnes qui ont participé à la rédaction des transcriptions originales, notamment Paloma Acevedo, Carlos Asenjo, Sebastian Bauhoff, Bradley Chen, Changcheng Song, Jane Zhang et Shufang Zhang. Nous tenons également à exprimer notre reconnaissance à Kristine Cronin pour la qualité de son travail d'assistance de recherche, à Marco Guzman et Martin Ruegenberg pour leurs illustrations ainsi qu'à Cindy A. Fisher, Fiona Mackintosh et Stuart K. Tucker pour leur travail éditorial lors de la rédaction de ce manuel.

Nous voudrions aussi reconnaître le soutien et l'engagement en faveur de ce type de travaux au sein de la Banque mondiale, notamment de la part d'Ariel Fiszbein, Arianna Legovini et Martin Ravallion.

Enfin, nous voudrions remercier l'ensemble des participants aux ateliers qui se sont tenus à Mexico, New Delhi, Cuernavaca, Ankara, Buenos Aires, Paipa, Fortaleza, Sofia, Managua, Madrid, Washington, Manille, Pretoria, Tunis, Lima, Amman, Beijing, Sarajevo, San Salvador, Katmandu, Rio de Janeiro, Accra, Séoul ainsi qu'au Caire et au Cap. Leur intérêt, leurs questions pertinentes et leur enthousiasme nous ont peu à peu appris ce que les décideurs recherchaient en matière d'évaluations d'impact. Nous espérons que ce manuel reflète leurs idées.

Références

Baker, Judy. 2000. *Evaluating the Impact of Development Projects on Poverty*. Washington DC : Banque mondiale.

Duflo Esther, Rachel Glennerster et Michael Kremer. 2007. « Using Randomization in Development Economics Research: A Toolkit. » Document de travail du CEPR no 6059. Center for Economic Policy Research, Londres, Royaume-Uni.

Duflo Esther et Michael Kremer. 2008. « Use of Randomization in the Evaluation of Development Effectiveness. » In *Evaluating Development Effectiveness*, vol. 7. Washington, DC : Banque mondiale.

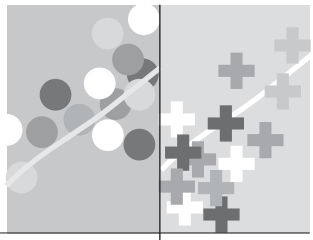
Khandker, Shahidur R., Gayatri B. Koolwal et Hussain Samad. 2009. *Handbook on Quantitative Methods of Program Evaluation*. Washington DC : Banque mondiale.

Leeuw, Frans et Jos Vaessen. 2009. *Impact Evaluations and Development*. *NONIE Guidance on Impact Evaluation*. Washington DC : NONIE et Banque mondiale.

Ravallion, Martin. 2001. « The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation. » *Étude économique de la Banque mondiale* 15 (1) : 115–40.

———. 2008. « Evaluating Anti-Poverty Programs. » In *Handbook of Development Economics*, vol, 4., éd. Paul Schultz et John Strauss. Amsterdam : Hollande-Septentrionale

———. 2009. « Evaluation in the Practice of Development. » *World Bank Research Observer* 24 (1) : 29–53.



Partie 1

INTRODUCTION À L'ÉVALUATION D'IMPACT

La première partie de l'ouvrage présente un aperçu général de l'évaluation d'impact. Dans le chapitre 1, nous exposons les raisons pour lesquelles l'évaluation d'impact est importante et comment elle s'inscrit dans le cadre de la formulation de politiques fondée sur les preuves. Nous comparons l'évaluation d'impact avec d'autres méthodes d'évaluation courantes telles que le suivi et les évaluations de processus. Nous présentons aussi différents types d'évaluation d'impact, comme les évaluations prospectives et rétrospectives ou encore les études d'efficacité pilotes et les études d'efficacité à l'échelle.

Dans le chapitre 2, nous traitons de la formulation de questions d'évaluation et d'hypothèses utiles aux décisions de politiques. Ces questions et hypothèses jouent un rôle primordial, car elles déterminent l'objectif de l'évaluation.

CHAPITRE 1

Pourquoi évaluer ?

Les politiques et les programmes de développement sont généralement conçus pour améliorer des indicateurs de résultats, par exemple pour augmenter les revenus, faciliter l'apprentissage ou faire reculer la maladie. Savoir si les changements espérés se sont effectivement produits est une question de politique publique importante, et pourtant rarement considérée. Le plus souvent, les responsables de programme et les décideurs mettent l'accent sur le contrôle et la mesure des intrants et des produits immédiats (par exemple le montant d'argent dépensé et la quantité de livres distribués) plutôt que sur la question de savoir si les programmes ont atteint leurs objectifs en termes d'amélioration du bien-être des bénéficiaires.

Élaboration des politiques fondée sur les preuves

Les évaluations d'impact s'inscrivent dans la vaste tendance de l'*élaboration de politiques fondée sur les preuves*. Cette tendance internationale croissante accorde une attention particulière aux extrants et aux résultats au-delà des seuls intrants. Dans ce cadre, la mise en œuvre de politiques publiques est en train de se transformer, comme l'illustrent les Objectifs de développement pour le Millénaire ou les initiatives de paiement à la performance des prestataires de services. Une attention particulière sur les résultats est utile pour fixer des objectifs nationaux et internationaux et en garantir le suivi. Elle est également de plus en plus requise des responsables de programme pour davantage les responsabiliser, justifier les allocations budgétaires et orienter les décisions de politique publique.

Le suivi et l'évaluation sont au cœur de l'élaboration de politiques fondée sur les preuves. Ils constituent en effet les outils élémentaires que les diverses parties pre-

nantes peuvent utiliser pour vérifier et améliorer la qualité, l'efficience et l'efficacité des programmes à différents stades de mise en œuvre. Autrement dit, le suivi et l'évaluation permettent de se focaliser sur les résultats. Tant les officiels des gouvernements que les acteurs extérieurs peuvent bénéficier de l'usage du suivi et de l'évaluation. Dans un ministère ou un organisme public, les fonctionnaires ont souvent besoin de prouver à leurs supérieurs que les programmes produisent des résultats afin d'obtenir les allocations budgétaires nécessaires à la poursuite ou l'amplification desdits programmes. Au plan national, les ministères sectoriels sont parfois en concurrence directe pour obtenir des fonds du ministère des Finances. Les gouvernements eux-mêmes doivent convaincre leurs électeurs que les investissements qu'ils ont choisis portent leurs fruits. En ce sens, les informations et des preuves solides constituent un moyen de sensibiliser le public et d'encourager la responsabilisation du gouvernement. L'information produite par les systèmes de suivi et d'évaluation peut être régulièrement mise à la disposition des citoyens pour les informer des résultats des programmes publics, renforçant ainsi les pratiques de transparence et de responsabilisation des gouvernements.

Dans un contexte où les décideurs et la société civile demandent des résultats et exigent que les responsables de programme rendent compte de la performance de leurs interventions, l'évaluation d'impact fournit des données solides et fiables qui indiquent si un programme donné a effectivement atteint les objectifs espérés. Au plan international, les évaluations d'impact jouent également un rôle crucial dans la mesure où elles permettent de mieux cerner l'efficacité des programmes de développement en mettant en évidence ce qui fonctionne et ce qui ne fonctionne pas en matière de réduction de la pauvreté et d'amélioration du bien-être des populations.

En résumé, une évaluation d'impact permet d'identifier les changements du bien-être des individus qui peuvent être *attribués* à un projet, un programme ou une politique particulière. Le concept de l'attribution est au cœur des évaluations d'impact. De ce fait, le principal défi d'une évaluation d'impact est d'identifier la *relation causale* entre un projet, un programme ou une politique et les résultats à l'étude.

Comme nous le verrons ci-dessous, les évaluations d'impact visent généralement à estimer l'impact *moyen* d'un programme sur le bien-être des bénéficiaires. Par exemple, la mise en œuvre d'un nouveau programme scolaire entraîne-t-elle de meilleurs résultats aux examens ? Un programme d'accès à l'eau potable permet-il d'améliorer les indicateurs de santé des bénéficiaires ? Un programme de formation des jeunes permet-il d'encourager l'entrepreneuriat et d'augmenter les revenus ? En outre, si l'évaluation d'impact est fondée sur un échantillon suffisamment grand, les impacts du programme peuvent également être comparés entre divers sous-groupes de bénéficiaires. Par exemple, le lancement d'un nouveau programme scolaire entraîne-t-il de meilleurs résultats aux examens tant pour les filles que pour les garçons ? Les évaluations d'impact peuvent aussi permettre de tester explicitement des options alternatives de concevoir des programmes. Par exemple, une évaluation peut comparer la performance d'un programme de formation visant à renforcer les connaissances financières des bénéficiaires par contraste avec une campagne de promotion ayant les mêmes objectifs. Dans chaque cas, l'évaluation d'impact donne des informations sur l'impact *général* du programme contrairement aux études de cas ou aux anecdotes fournissant des informations parcellaires qui ne reflètent pas forcément l'impact

général du programme. Dans ce sens, une évaluation bien conçue et correctement mise en œuvre permet d'obtenir des données convaincantes et exhaustives qui peuvent être utilisées pour orienter les décisions de politique et informer l'opinion publique. L'encadré 1.1 illustre comment l'évaluation d'impact a contribué au débat sur l'élargissement d'un programme de transferts monétaires conditionnels au Mexique[1].

Encadré 1.1 : Evaluation et durabilité politique

Le programme de transferts monétaires conditionnels Progresa/Oportunidades au Mexique

Dans les années 90, les autorités mexicaines lancent un programme innovateur de transferts monétaires conditionnels (TMC) baptisé Progresa. Les objectifs de ce programme sont de fournir aux ménages pauvres un soutien financier à court terme et d'encourager l'investissement dans le capital humain des enfants, essentiellement en octroyant aux mères des ménages pauvres une allocation monétaire à condition que leurs enfants soient scolarisés et effectuent régulièrement des examens de santé.

Dès le départ, le gouvernement met l'accent sur l'importance du suivi et de l'évaluation du programme. Les responsables chargent un groupe de chercheurs de concevoir une évaluation d'impact et de l'appliquer au programme au fur et à mesure de son expansion aux diverses communautés concernées.

Les élections présidentielles de 2000 conduisent à un changement du parti au pouvoir. En 2001, les évaluateurs externes du programme Progresa présentent leurs conclusions à la nouvelle administration. Les résultats s'avèrent impressionnants : le programme est bien ciblé aux populations pauvres et engendre des changements prometteurs en matière de capital humain. Schultz (2004) montre que le programme augmente le taux de scolarisation et allonge la durée de la scolarisation de 0,7 année en moyenne. De plus, selon Gertler (2004), l'incidence des maladies recule de 23 % chez les enfants tandis que pour les adultes le nombre de jours de travail perdus pour cause de maladie ou d'incapacité diminue de 19 %. Au niveau nutritionnel, Behrman and Hoddinott (2001) montrent que le programme réduit le retard de croissance d'environ 1 cm par an chez les enfants se situant dans la tranche d'âge critique de 12 à 36 mois.

Ces résultats permettent d'alimenter un dialogue politique fondé sur des preuves et incitent la nouvelle administration à maintenir le programme. Le gouvernement élargit même la couverture du programme, y intégrant l'octroi de bourses pour le collège et le lycée ainsi que des programmes d'amélioration de la santé des adolescents. Parallèlement, les résultats de l'évaluation conduisent à modifier d'autres programmes d'assistance sociale. Par exemple, le programme de subvention de « tortilla », couteux et moins bien ciblé, est revu à la baisse.

Le succès de l'évaluation du programme Progresa contribue aussi au rapide développement des TMC à travers le monde ainsi qu'à l'adoption d'une loi exigeant que tous les projets sociaux fassent l'objet d'une évaluation au Mexique.

Source : Behrman et Hoddinott 2001 ; Gertler 2004 ; Fiszbein et Schady 2009 ; Levy et Rodriguez 2005 ; Schultz 2004 ; Skoufias et McClafferty 2001.

L'encadré 1.2 montre, quant à lui, comment l'évaluation d'impact a mené à l'amélioration de l'allocation des ressources du gouvernement indonésien en permettant d'identifier les politiques les plus efficaces pour réduire le taux de fécondité.

Encadré 1.2 : L'évaluation au service d'une meilleure allocation des ressources

Planification familiale et fécondité en Indonésie

Dans les années 70, la politique de planification familiale indonésienne acquiert une reconnaissance internationale pour ses succès en matière de baisse du taux de fécondité. Cette renommée provient de deux phénomènes parallèles : 1) le taux de fertilité diminue de 22 % entre 1970 et 1980, de 25 % entre 1981 et 1990, et d'un peu moins entre 1991 et 1994 ; et 2) au cours de la même période, le Gouvernement indonésien augmente fortement les ressources consacrées à la planification familiale (notamment les subventions pour les contraceptifs). Ces deux phénomènes étant concomitants, nombreux sont ceux qui concluent que la hausse des investissements dans la planification familiale provoque la baisse de la fécondité.

Sceptiques, des chercheurs se demandent si les programmes de planification familiale sont réellement à l'origine de la baisse de la fécondité. Contre toute attente, ils découvrent que ces programmes n'ont qu'un impact limité sur la fécondité et concluent que cette baisse s'explique essentiellement par un changement du statut des femmes. Les chercheurs font remarquer qu'avant le lancement du programme de planification familiale, très peu de femmes en âge de procréer avaient terminé le cycle primaire. En parallèle à la planification familiale, le gouvernement avait lancé un vaste programme de scolarisation des filles, impliquant qu'elles sont plus nombreuses à avoir acquis une éducation au moment d'atteindre l'âge de procréation. Parallèlement, la croissance économique et l'offre accrue d'emplois engendrées par le boom pétrolier entraînent une augmentation du nombre de femmes instruites dans la population active. L'augmentation du temps de travail des femmes s'accompagne d'une hausse de l'usage des contraceptifs. In fine, l'augmentation des revenus et l'autonomisation des femmes expliquent 70 % de la baisse de fécondité observée, soit plus que les investissements dans les programmes de planification familiale.

Ces résultats permettent aux décideurs politiques de réorienter l'allocation des ressources en connaissance de cause : les subventions en faveur de la contraception sont réduites au profit des programmes encourageant la scolarisation des femmes. Les objectifs finaux des deux types de programmes sont certes les mêmes, mais les études d'évaluation ont mis en évidence que, dans le cas de l'Indonésie, les investissements dans l'éducation sont plus efficaces que les investissements dans la planification familiale pour réduire la fécondité.

Source : Gertler et Molyneaux 1994, 2000.

Qu'est-ce que l'évaluation d'impact ?

L'évaluation d'impact fait partie d'une large gamme de méthodes complémentaires contribuant à l'élaboration de politiques fondée sur des preuves. Le présent ouvrage est consacré aux méthodes d'évaluation d'impact quantitatives ; nous allons cependant commencer par les placer dans le cadre plus général de la gestion publique axée sur les résultats, qui comprend aussi le suivi et d'autres types d'évaluation.

Le suivi est un processus continu qui consiste à surveiller le déroulement d'un programme et qui s'appuie sur les données collectées pour améliorer la mise en œuvre du programme, sa gestion et les décisions quotidiennes le concernant. Ce processus s'appuie essentiellement sur les données administratives pour comparer la performance du programme aux résultats espérés, comparer les programmes entre eux et analyser des tendances à travers le temps. Le suivi se concentre généralement sur les intrants, les activités et les extrants, ainsi qu'occasionnellement les résultats, comme les progrès vers les objectifs de développement nationaux.

L'évaluation est une appréciation périodique et objective de projets, programmes ou politiques prévus, en cours de réalisation ou achevés. Les évaluations permettent de répondre à des questions précises liées à la conception, la mise en œuvre ou les résultats des programmes. Contrairement au suivi, qui est continu, les évaluations sont périodiques et effectuées à un moment donné, généralement par des spécialistes extérieurs au programme. La conception, la méthodologie et le coût des évaluations varient fortement en fonction du type de question à laquelle elles répondent. D'une manière générale, les évaluations s'attèlent à trois types de questions (Imas et Rist, 2009) :

- *Les questions descriptives :* à ce niveau, l'évaluation vise à montrer ce qui se passe, décrire les processus, les conditions qui prévalent, les relations organisationnelles et les points de vue des diverses parties prenantes au programme.
- *Les questions normatives :* l'évaluation compare ce qui se passe à ce qui devrait se passer ; elle consiste à étudier les activités et à estimer si les objectifs ont été atteints ou non. Les questions normatives peuvent concerner les intrants, les activités et les extrants.
- *Les questions de cause à effet :* l'évaluation se concentre sur les résultats et cherche à déterminer dans quelle mesure l'intervention entraine des changements des résultats.

Les *évaluations d'impact* sont un type particulier d'évaluation qui porte sur les questions de cause à effet. Contrairement aux évaluations générales qui peuvent répondre à plusieurs types de questions, les évaluations d'impact sont structurées autour d'un type particulier de question : *quel est l'impact (ou l'effet causal) d'un programme sur un résultat donné ?* La dimension causale est primordiale. Nous nous intéressons ici à *l'impact* du programme, à savoir les changements des résultats causés directement par celui-ci. L'évaluation d'impact vise ainsi à déterminer quels changements peuvent être *attribués directement et exclusivement au programme*.

L'évaluation d'impact se distingue par sa focalisation sur la causalité et l'attribution des changements, deux concepts qui définissent aussi l'approche méthodologique. Pour pouvoir estimer l'effet causal ou l'impact d'un programme sur les résultats, la méthode choisie doit permettre de définir le contrefactuel, c'est-à-dire le résultat qui aurait été obtenu pour un groupe de bénéficiaires si le programme n'avait pas existé. Dans les faits, l'évaluation d'impact nécessite de trouver un groupe de comparaison pour estimer les résultats qu'auraient connus les participants à un programme si ledit programme n'avait pas existé. La partie 2 du manuel est consacrée aux principales méthodes utilisées pour constituer des groupes de comparaison adéquats.

Concept clé :
La question fondamentale de l'évaluation d'impact peut être formulée ainsi : quel est l'impact (ou l'effet causal) d'un programme sur un résultat donné ?

La question fondamentale de l'évaluation d'impact (à savoir *quel est l'impact ou l'effet causal d'un programme sur un résultat donné ?*) peut être appliquée à de nombreux contextes. Par exemple, quel est l'impact de l'octroi de bourses sur la scolarisation et les résultats académiques ? Quel est l'impact sur l'accès aux soins de santé de la sous-traitance de ces services à des prestataires privés ? Quel est l'impact de remplacer des sols en terre battue par des sols en ciment sur la santé des enfants ? L'amélioration de l'état des routes permet-elle un meilleur accès au marché du travail et une augmentation des revenus des ménages, et si tel est le cas, à quel degré ? La taille des classes a-t-elle un impact sur les résultats scolaires des étudiants et, si oui, dans quelle mesure ? Quelle est l'efficacité des campagnes de publipostage relativement à des formations pratiques lorsqu'il s'agit d'augmenter l'utilisation des moustiquaires dans les zones affectées par le paludisme ?

L'évaluation d'impact pour les décisions politiques

Les évaluations d'impact permettent d'éclairer les responsables politiques lorsqu'ils prennent plusieurs types de décisions : interruption des programmes inefficaces, expansion des interventions éprouvées, ajustement des bénéfices, sélection entre plusieurs options de conception de programmes. Pour être efficaces, les évaluations doivent être utilisées de manière sélective pour répondre aux questions de politique publique les plus importantes. Elles s'avèrent particulièrement utiles dans le cadre des programmes pilotes qui éprouvent des approches nouvelles et prometteuses n'ayant pas encore fait leurs preuves. L'évaluation du programme Progresa/Oportunidades au Mexique présentée dans l'encadré 1.1 a été très influente, non seulement du fait du caractère novateur du programme, mais aussi parce qu'elle a fourni des preuves fiables et solides qui ne pouvaient être ignorées dans les prises de décision ultérieures. L'adoption et l'élargissement du programme ont été largement influencés par les résultats de l'évaluation. Aujourd'hui, le programme Oportunidades bénéficie près d'un Mexicain sur quatre et forme le cœur de la stratégie de lutte contre la pauvreté du Mexique.

Les évaluations d'impact peuvent servir à analyser différents types de questions de politique publique. Dans leur forme élémentaire, elles permettent de tester l'efficacité d'un programme donné. Autrement dit, elles répondent à la question suivante : *un programme donné est-il efficace en comparaison à l'absence de ce programme ?* Comme nous

le verrons dans la partie 2, ce type d'évaluation d'impact estime l'efficacité du programme en comparant un groupe de traitement qui a bénéficié d'un projet, d'un programme ou d'une politique avec un groupe de comparaison qui n'y a pas participé.

Outre la réponse à la question fondamentale d'évaluation, les évaluations peuvent aussi servir à mesurer l'efficacité de diverses options de mise en œuvre d'un programme, autrement dit de répondre à la question suivante : *lorsqu'un programme peut être mis en œuvre de plusieurs manières, y en a-t-il une qui est plus efficace que les autres ?* Dans ce type d'évaluation, deux ou plusieurs options de concevoir un même programme sont comparées, de façon à déterminer le meilleur moyen d'atteindre un objectif particulier. Ces différentes options sont souvent appelées « branches de traitement ». Par exemple, quand la taille du bénéfice nécessaire pour rendre un programme efficace n'est pas connue (20 heures ou 80 heures de formation ?), les évaluations d'impact peuvent permettre d'estimer l'impact relatif de traitements d'intensités différentes (voir exemple de l'encadré 1.3). Les évaluations d'impact destinées à tester différentes options de traitement incluent généralement un groupe de traitement par branche, ainsi qu'un groupe de comparaison « pur » qui, lui, n'est pas soumis à l'intervention. Les évaluations d'impact peuvent être utiles pour tester des innovations ou des alternatives de mise en œuvre d'un programme. Par exemple, il est possible de mettre en œuvre plusieurs formes de campagnes de sensibilisation afin de déterminer l'approche la plus efficace : un groupe est sélectionné pour recevoir une campagne de publipostage tandis que d'autres groupes recevront des visites à domicile.

Encadré 1.3 : L'évaluation au service d'une meilleure conception des programmes

Malnutrition et développement cognitif en Colombie

Au début des années 70, la Human Ecology Research Station lance, en collaboration avec le ministère colombien de l'Éducation, un programme destiné à lutter contre la malnutrition infantile dans la ville de Cali en fournissant des soins de santé, des activités d'éducation, de la nourriture et des compléments alimentaires. Dans le cadre de la phase pilote, une équipe d'évaluateurs est chargée de déterminer 1) le temps nécessaire pour qu'un tel programme réduise la malnutrition chez les enfants d'âge préscolaire issus de familles à faibles revenus et 2) si les interventions peuvent aussi permettre des améliorations sur le plan du développement cognitif.

Le programme est ouvert à toutes les familles éligibles, mais durant la phase pilote les évaluateurs comparent des groupes similaires d'enfants ayant reçu le traitement sur des durées différentes. Les évaluateurs commencent par sélectionner un groupe cible de 333 enfants souffrant de malnutrition. Ces enfants sont ensuite classés en 20 secteurs selon leur lieu d'habitation, et chaque secteur se voit assigné, de manière aléatoire, à l'un des quatre groupes de traitement. La seule différence entre les groupes est le moment auquel ils commencent à recevoir le traitement et, de ce fait, la durée pendant laquelle ils participent au programme. Le groupe 4 commence le premier. Il est donc exposé au traitement le plus longtemps. Suivent les groupes 3, puis 2, puis 1. Le traitement consiste en six heures quotidiennes de soins de santé

(suite)

Encadré 1.3 *suite*

et d'activités éducatives, et en la provision de nourriture et de compléments alimentaires. Au cours de la période de mise en œuvre du programme, les évaluateurs procèdent à intervalles réguliers à des tests cognitifs pour suivre les progrès des enfants de chacun des quatre groupes.

Les évaluateurs découvrent que les enfants ayant participé au programme le plus longtemps sont ceux qui enregistrent les améliorations cognitives les plus importantes. Au test Stanford-Binet, qui évalue la différence entre l'âge mental et l'âge chronologique, les enfants du groupe 4 ont une différence moyenne de -5 mois, contre -15 mois pour le groupe 1.

Cet exemple montre que les responsables du programme et les décideurs politiques peuvent recourir à l'évaluation de plusieurs branches de traitement pour déterminer l'alternative la plus efficace.

Source : McKay et al. 1978.

Décider quand évaluer

Tous les programmes ne nécessitent pas une évaluation d'impact. Ces évaluations peuvent être coûteuses et le budget prévu pour les évaluations doit être utilisé de manière stratégique. Si vous lancez un nouveau programme ou si vous envisagez l'élargissement d'un programme en vigueur et que vous vous demandez si une évaluation d'impact est nécessaire, quelques questions peuvent vous aider à trancher.

La première question à se poser est la suivante : *quels sont les enjeux de ce programme ?* La réponse dépend à la fois des montants engagés et du nombre de personnes qui sont ou seront touchées par le programme. D'où les questions suivantes : *Le programme nécessite-t-il ou nécessitera-t-il une grande partie du budget disponible ?* et *Le programme touche-t-il ou touchera-t-il un nombre important de personnes ?* S'il se trouve que le programme ne consommera pas un budget important ou qu'il ne concernera qu'un nombre limité de personnes, une évaluation n'est pas forcement utile. Par exemple, pour un programme d'aide et de soutien délivrés par des volontaires à des patients hospitalisés, le budget et le nombre de bénéficiaires peuvent être tels qu'une évaluation d'impact ne se justifie pas. À l'inverse, pour une réforme des salaires de l'ensemble des enseignants du primaire d'un pays, les enjeux sont nettement plus importants.

Si vous considérez que les enjeux sont de taille, la question qui se pose alors est de savoir s'il existe des données permettant de montrer que le programme donne des résultats. En particulier, avez-vous une idée de l'ampleur de l'impact du programme ? Existe-t-il des données concernant un pays et un contexte similaires ? S'il n'existe aucune information sur l'impact potentiel du programme envisagé, vous pouvez commencer par une phase pilote avec une évaluation d'impact. En revanche, si vous disposez déjà de données sur une situation similaire, l'évaluation d'impact ne se jus-

tifiera probablement que si elle permet de répondre à une nouvelle question de politique importante. Ce sera par exemple le cas si votre programme contient des innovations importantes qui n'ont encore jamais été éprouvées.

Pour justifier la mobilisation des ressources techniques et financières nécessaires à la réalisation d'une évaluation d'impact de qualité, le programme à évaluer doit être :

- *Novateur*. Il permet de tester une nouvelle approche prometteuse.
- *Reproductible*. Le programme peut être élargi et reproduit dans un autre contexte.
- *Stratégiquement pertinent*. Le programme est une initiative phare ; il nécessite des ressources importantes ; il couvre ou couvrira un grand nombre de bénéficiaires ; ou encore il permettrait de faire des économies importantes.
- *Non testé auparavant*. L'efficacité du programme est méconnue soit au niveau international, soit dans un contexte particulier.
- *Influent*. Les résultats du programme permettront d'orienter des décisions de politique clés.

Analyse du rapport coût-efficacité

Concept clé :
L'analyse coût-bénéfice permet d'estimer les bénéfices totaux espérés du programme par rapport aux coûts totaux prévus.

Lorsque les résultats de l'évaluation d'impact sont disponibles, ils peuvent être combinés aux données sur les coûts du programme pour traiter deux autres types de questions. Tout d'abord, pour les évaluations d'impact les plus élémentaires, le fait de prendre en compte les coûts permettra de réaliser une analyse coût-bénéfice et de répondre à la question suivante : *quel est le rapport coût-bénéfice d'un programme donné ?* L'analyse coût-bénéfice permet d'estimer les bénéfices totaux espérés du programme par rapport à ses coûts totaux. L'objectif est de déterminer l'ensemble des coûts et des bénéfices monétaires d'un programme et de voir ainsi si les bénéfices sont supérieurs aux coûts.

Dans un monde parfait, une analyse coût-bénéfice fondée sur les résultats concrets de l'évaluation d'impact pourrait être réalisée non seulement pour un programme donné, mais aussi pour toute une série de programmes ou d'alternatives de conception d'un même programme. Les décideurs politiques seraient ainsi en mesure de choisir en toute certitude le programme ou l'approche présentant le meilleur rapport coût-bénéfice pour atteindre un objectif donné. Lorsqu'une évaluation d'impact porte sur des alternatives de mise en œuvre d'un même programme, la prise en compte des informations de coûts permet de répondre à une seconde question : *quels sont les rapports coût-efficacité des diverses approches ?* L'analyse coût-efficacité compare la performance relative de deux ou plusieurs programmes ou alternatives de conception d'un programme à atteindre un même résultat.

Concept clé :
L'analyse du rapport coût-efficacité compare la performance relative de deux ou plusieurs programmes ou alternatives de conception d'un programme à atteindre un même résultat.

Qu'il s'agisse d'analyse coût-bénéfice ou de rapport coût-efficacité, l'évaluation d'impact permet d'estimer les bénéfices et l'efficacité, tandis que les informations de coûts sont fournies par l'analyse des coûts. Le présent manuel porte sur l'éva-

luation d'impact et ne traite pas en détail des questions relatives à la collecte des informations sur les coûts ou à l'analyse coût-bénéfice[2]. Il est toutefois primordial de disposer des informations relatives aux coûts du projet, du programme ou de la politique qui fait l'objet de l'évaluation. Lorsque des informations sur l'impact et les coûts de divers programmes sont disponibles, le rapport coût-efficacité permet de déterminer les investissements les plus rentables et d'orienter ainsi les décisions des responsables. L'encadré 1.4 illustre comment les évaluations d'impact peuvent servir à déterminer les programmes les plus rentables et à mieux allouer les ressources.

Encadré 1.4 : Évaluation du rapport coût-efficacité

Comparaison de stratégies pour augmenter la fréquentation scolaire au Kenya

En évaluant plusieurs programmes dans un même contexte, il est possible de comparer le rapport coût-efficacité de différentes approches visant à améliorer un résultat donné, par exemple la fréquentation scolaire. Au Kenya, l'organisation non gouvernementale International Child Support Africa (ICS Africa) met en œuvre toute une série d'interventions en milieu scolaire qui comprennent un traitement contre les vers intestinaux ainsi que la fourniture gratuite d'uniformes et de repas scolaires. Chacune de ces interventions fait l'objet d'une évaluation aléatoire et d'une analyse coût-bénéfice ; les comparaisons entre les différentes interventions génèrent des informations intéressantes sur la meilleure manière d'augmenter la fréquentation scolaire.

Un programme proposant des médicaments contre les vers intestinaux aux enfants scolarisés entraîne une hausse de la fréquentation de l'ordre de 0,14 an par enfant traité, pour un coût estimé à 0,49 dollar par enfant. Ceci représente environ 3,50 dollars par année supplémentaire de scolarisation, compte tenu des externalités sur les enfants et les adultes ne fréquentant pas l'école, mais vivant dans les communautés qui bénéficient indirectement d'une diminution de la transmission des vers.

Une seconde intervention, le Child Sponsorship Program, permet de réduire le coût de scolarisation en fournissant des uniformes scolaires aux enfants de sept écoles sélectionnées de manière aléatoire. Le taux d'abandon chute dans les écoles bénéficiant de l'intervention ; après cinq ans, il est estimé que le programme a permis d'augmenter le nombre d'années de scolarisation de 17 % en moyenne. Toutefois, même dans l'hypothèse la plus optimiste, le coût de cette augmentation de la fréquentation scolaire par la fourniture d'un uniforme scolaire ressort à environ 99 dollars par année supplémentaire de scolarisation.

Enfin, un programme consistant à fournir gratuitement un petit déjeuner aux enfants de 25 écoles maternelles sélectionnées aléatoirement entraîne une augmentation de 30 % de la fréquentation scolaire pour un coût estimé à 36 dollars par année de scolarisation supplémentaire. Les résultats aux examens augmentent aussi, d'environ 0,4 au niveau de l'écart-type, lorsque l'enseignant est bien formé avant le lancement du programme.

Bien que des interventions similaires puissent cibler différents résultats (par exemple une amélioration de l'état de santé grâce aux vermifuges ou de meilleurs résultats scolaires parallèlement à la hausse de la fréquentation scolaire), la comparaison de diverses évaluations menées dans un même contexte permet de déterminer les programmes qui ont atteint l'objectif visé au meilleur coût

Source : Kremer et Miguel 2004 ; Kremer, Moulin et Namunyu 2003 ; Poverty Action Lab 2005 ; Vermeersch et Kremer 2005.

Évaluation prospective et évaluation rétrospective

Les évaluations d'impact peuvent être regroupées en deux catégories : les évaluations prospectives et les évaluations rétrospectives. Les évaluations prospectives sont prévues dès la conception du programme et font partie intégrante de sa mise en œuvre. Les données de l'enquête de base (ou enquête de référence) sont collectées avant la mise en place du programme tant pour le groupe de traitement que pour le groupe de comparaison. Les évaluations rétrospectives portent, quant à elles, sur l'impact du programme après la mise en œuvre de celui-ci, les groupes de traitement et de comparaison étant définis *ex-post*.

En général, les évaluations d'impact prospectives donnent des résultats plus solides et plus fiables, et ce pour trois raisons.

Concept clé :
Les évaluations prospectives sont élaborées dès la conception du programme et font partie intégrante de la mise en œuvre du programme.

En premier lieu, la collecte préalable de données de base (ou enquête de référence) permet d'assurer la mesure des résultats à l'étude. Les données de base fournissent des informations sur les bénéficiaires et les groupes de comparaison avant la mise en œuvre du programme et sont donc primordiales pour connaître la situation avant le programme. Une enquête de référence couvrant le groupe de traitement et le groupe de comparaison peut être analysée pour vérifier que ces groupes sont bien similaires. Elle peut par ailleurs permettre d'évaluer l'efficacité du ciblage, autrement dit d'établir si le programme touche effectivement les bénéficiaires visés.

En deuxième lieu, la définition de mesures pour juger du succès d'un programme dès sa conception permet d'axer non seulement l'évaluation, mais aussi le programme sur les résultats espérés. Comme nous le verrons, les évaluations d'impact découlent d'une théorie du changement ou chaîne de résultats. La conception d'une évaluation d'impact contribue à mieux définir les objectifs du programme, en particulier parce qu'elle exige d'établir des mesures pour juger de l'efficacité du programme. Les décideurs doivent définir des questions et des objectifs d'évaluation clairs de manière à ce que les résultats soient des plus pertinents. Le soutien total des décideurs est en effet une condition préalable de la réalisation d'une évaluation ; une évaluation d'impact ne doit pas être engagée si les décideurs ne sont pas convaincus de sa légitimité et de son importance pour éclairer les décisions futures.

La troisième raison est la plus importante : dans une évaluation prospective, les groupes de traitement et de comparaison sont définis avant l'entrée en vigueur du programme. Comme nous l'expliquerons plus en détail dans les chapitres suivants, de nombreuses options existent pour réaliser des évaluations valides si celles-ci sont prévues dès le départ et informées par la mise en œuvre du projet. Comme nous le montrons dans les parties 2 et 3, si l'évaluation prospective est bien conçue, une estimation valide du contrefactuel est possible pour tout programme suivant des règles d'assignation claires et transparentes. En bref, l'évaluation prospective a de meilleures chances de générer une estimation valide du contrefactuel. Différentes manières d'élaborer un contrefactuel valide peuvent être considérées dès la conception du programme, et la méthodologie d'évaluation d'impact peut ainsi être totalement alignée sur les règles opérationnelles du programme, son déroulement ou son élargissement.

À l'inverse, dans les évaluations rétrospectives, l'évaluateur dispose souvent de si peu d'informations qu'il lui est difficile de déterminer si le programme a été mis en œuvre avec succès et si les participants y ont effectivement pris part. En effet, pour de nombreux programmes, il n'existe pas de données de base lorsque l'évaluation n'est pas intégrée au projet dès le départ. Une fois le programme lancé, il est trop tard pour collecter les données de base nécessaires.

L'évaluation de programmes mis en œuvre par le passé ne peut se faire que par une évaluation rétrospective se fondant sur des données existantes. Dans ce cas, il est généralement beaucoup plus difficile de définir un contrefactuel valide. L'évaluation dépend de l'application des règles opérationnelles précises de distribution des bénéfices. Elle est également tributaire de la disponibilité des données pour les groupes de traitement et de comparaison tant avant qu'après l'entrée en vigueur du programme. Par conséquent, la faisabilité d'une évaluation rétrospective dépend du contexte et n'est jamais garantie. Même lorsqu'elle est faisable, l'évaluation rétrospective repose souvent sur des méthodes quasi-expérimentales et des hypothèses plus fortes ; les résultats sont donc plus discutables.

Études d'efficacité pilotes et études d'efficacité à l'échelle

Le principal rôle d'une évaluation d'impact est de produire des preuves quant à l'efficacité d'un programme à l'usage des décideurs politiques, des responsables de programme, de la société civile ainsi que de toute autre partie prenante. Les résultats d'une évaluation d'impact sont particulièrement utiles lorsque les conclusions peuvent être appliquées à une population plus large. La question de la généralisation des conclusions (ou « validité externe » dans le jargon des méthodes de recherche) est centrale pour les décideurs, car elle permet d'établir si les résultats obtenus par l'évaluation peuvent s'appliquer à des groupes autres que ceux qui ont été étudiés, ce qui est primordial si un élargissement du programme est envisagé.

Les premières évaluations d'impact de programmes de développement constituaient souvent des *études d'efficacité pilotes* menées dans des conditions très particulières. Malheureusement, les résultats de ces études ne pouvaient que rarement être généralisés au-delà du contexte de l'évaluation. Les études d'efficacité pilotes sont généralement réalisées dans des conditions très particulières et avec une assistance technique importante tout au long de la mise en œuvre du programme. Ces études sont souvent mises en œuvre pour valider un concept ou tester la viabilité d'un nouveau programme. Si le programme ne génère pas l'impact prévu dans les conditions de l'étude, qui sont souvent bien maîtrisées, il a peu de chances de donner des résultats s'il est appliqué dans des conditions normales. Les études pilotes sont généralement de petite envergure et mises en œuvre dans des conditions étroite-

ment contrôlées ; l'impact qu'elles permettent de mettre en évidence risque de ne pas être représentatif de l'impact d'un projet similaire réalisé à plus grande échelle et dans des conditions normales. Par exemple, un projet pilote de mise en vigueur de nouveaux protocoles médicaux peut donner de bons résultats dans un hôpital doté d'excellents gestionnaires et d'un bon personnel médical, mais se révéler totalement inefficace dans un hôpital moyen, si les gestionnaires sont moins attentifs et les membres du personnel moins nombreux. En outre, le rapport coût-bénéfice sera différent, car les coûts fixes et les économies d'échelle risquent de ne pas être pris en compte dans l'étude pilote vu l'envergure limitée du projet. Les études pilotes peuvent certes être utiles pour tester une approche novatrice, mais leurs résultats ont généralement une validité externe limitée et ne reflètent pas toujours les conditions réelles auxquelles les décideurs sont habituellement confrontés.

À l'inverse, les *études d'efficacité à l'échelle* caractérisent les interventions qui ont lieu dans des conditions normales et sont mises en œuvre par des voies habituelles. Lorsque les études d'efficacité à l'échelle sont bien conçues et bien réalisées, les résultats peuvent être considérés comme valides aussi bien pour l'échantillon d'évaluation que pour d'autres bénéficiaires potentiels hors de l'échantillon. La validité externe est primordiale pour les décideurs, car c'est elle qui définit s'il sera possible ou non d'utiliser les résultats de l'évaluation pour juger de l'opportunité d'étendre le programme au-delà de l'échantillon d'évaluation.

Combiner les sources d'information pour évaluer tant le « pourquoi » que le « comment »

Les évaluations d'impact réalisées sans tenir compte de diverses sources d'informations sont vulnérables tant sur le plan technique que sur le plan de leur efficacité potentielle. Sans informations sur la nature et le contenu du programme permettant de replacer les résultats de l'évaluation dans leur contexte, les décideurs ne pourront pas déterminer les raisons pour lesquelles un résultat a été atteint et non un autre. Si les évaluations d'impact donnent des estimations relativement fiables des effets causaux pour un programme, elles ne sont généralement pas conçues pour permettre d'analyser les aspects relatifs à l'efficience de la mise en œuvre du programme. De plus, elles doivent être en adéquation avec la réalisation du programme et doivent, en conséquence, tenir compte de la manière, du moment et du lieu où le programme évalué est exécuté.

Des données qualitatives, des données de suivi ainsi que des évaluations de processus sont nécessaires pour documenter la mise en œuvre d'un programme de façon à éclairer et interpréter les résultats des évaluations d'impact. À cet égard, les évaluations d'impact et les autres outils d'évaluation sont complémentaires les uns des autres plutôt que concurrents.

Par exemple, les autorités d'une province peuvent décider le versement de primes aux cliniques rurales qui réussissent à augmenter le pourcentage des naissances ayant lieu en présence d'un professionnel de la santé. Si l'évaluation montre qu'aucun changement n'a été constaté au niveau du pourcentage des naissances en cliniques, plusieurs explications peuvent être avancées. Il est tout d'abord possible que le personnel des cliniques concernées n'ait pas été suffisamment informé des primes ou qu'il n'ait pas compris les règles du programme. Dans ce cas, les autorités provinciales peuvent lancer une campagne d'information et d'éducation à l'attention des centres de santé. Il se peut aussi qu'un manque d'équipement ou des coupures d'électricité aient empêché les cliniques d'admettre plus de patients. Dans ce cas, il peut s'avérer nécessaire de renforcer les équipements et d'améliorer l'approvisionnement en électricité. Enfin, les femmes enceintes en milieu rural peuvent être réticentes à accoucher en clinique et préférer, pour des raisons culturelles, accoucher chez elles, assistées d'une sage-femme. Si tel est le cas, il sera sans doute plus efficace de s'attaquer aux barrières auxquelles se heurtent les femmes que de distribuer des primes aux cliniques. Une bonne évaluation d'impact permettra aux autorités de déterminer si l'évolution du taux des naissances en présence d'un professionnel de la santé est le résultat ou non de la distribution des primes. Des travaux complémentaires seront toutefois nécessaires pour déterminer si le programme s'est déroulé comme prévu et quelles en sont les pièces manquantes. Dans notre exemple, les évaluateurs peuvent compléter leur étude d'impact en interrogeant le personnel de santé des cliniques pour évaluer leur connaissance du programme, en examinant les équipements dont disposent les cliniques, en menant des discussions de groupe avec des femmes enceintes pour comprendre leurs préférences et leurs réticences, et en examinant l'ensemble des données disponibles sur l'accès aux centres de santé en milieu rural.

Utiliser des données qualitatives

Les *données qualitatives* constituent un complément important aux évaluations d'impact quantitatives, car elles peuvent donner des indications additionnelles sur la performance d'un programme. Les évaluations qui combinent l'analyse quantitative et l'analyse qualitative sont dites à « méthodes mixtes » (Bamberger, Rao et Woolcock, 2010). Les études qualitatives ont recours à des groupes focaux et à des entrevues avec certains bénéficiaires et d'autres personnes susceptibles de fournir des informations (Rao et Woolcock, 2003). Bien que les points de vue et opinions issus de ces entretiens et des groupes focaux ne puissent être considérés comme représentatifs de l'opinion de l'ensemble des bénéficiaires du programme, ils sont particulièrement utiles au cours des trois phases de l'évaluation d'impact :

1. Lors de la conception de l'évaluation d'impact, les évaluateurs peuvent avoir recours à des groupes focaux et interroger des personnes clés pour élaborer des

hypothèses sur la manière et les raisons de la réussite du programme, le cas échéant, et clarifier les questions de recherche auxquelles il s'agira de répondre lors de l'évaluation d'impact quantitative.

2. Au stade intermédiaire, soit avant que les résultats de l'évaluation quantitative ne soient connus, l'analyse qualitative peut permettre de fournir aux décideurs un aperçu de l'évolution du programme.

3. Au stade de l'analyse, les évaluateurs peuvent recourir aux méthodes qualitatives pour replacer les données quantitatives dans leur contexte et trouver des explications, pour mieux étudier les cas particuliers de réussite ou d'échec, et pour formuler des explications systématiques de la performance du programme établie par les résultats quantitatifs. En ce sens, l'analyse qualitative peut contribuer à expliquer certains résultats observés au terme de l'analyse quantitative et permettre de mieux comprendre ce qui s'est passé dans le cadre du programme (Bamberger, Rao et Woolcock, 2010).

Utiliser des données de suivi et des évaluations de processus

Les *données de suivi* sont également particulièrement précieuses pour l'évaluation d'impact. Elles permettent en effet de recenser les participants au programme, de déterminer la chronologie de développement du programme ou la manière dont les ressources sont dépensées, ainsi que d'une manière plus générale de vérifier si les activités sont mises en œuvre comme prévu. Ces informations sont très importantes pour la réalisation de l'évaluation, pour s'assurer par exemple que les données de l'enquête de référence sont bien collectées avant l'entrée en vigueur du programme ou encore pour vérifier l'adhérence à l'assignation aux groupes de traitement et de comparaison. En outre, le système de suivi peut fournir des informations sur le coût de la mise en œuvre du programme, particulièrement utiles pour l'analyse coût-bénéfice.

Pour leur part, les *évaluations de processus* mettent l'accent sur l'exécution et le déroulement du programme et visent à vérifier que le processus est conforme aux prévisions initiales ; elles fournissent des informations sur son développement et son déroulement. Ces évaluations peuvent généralement être effectuées assez rapidement et à un coût raisonnable. Dans le cadre des projets pilotes et des phases initiales de programmes, elles peuvent constituer des sources d'informations intéressantes pour améliorer l'exécution du programme.

Notes

1. Voir Fiszbein et Schady, 2009, pour un aperçu des programmes de TMC et le rôle influent du programme Progresa/Oportunidades suite à l'évaluation d'impact dont il a fait l'objet

2. Pour de plus amples informations sur l'analyse coût-bénéfice, voir Belli et al. 2001 ; Boardman et al. 2001 ; Brent 1996 ; ou Zerbe et Dively 1994.

Références

Bamberger, Michael, Vijayendra Rao et Michael Woolcock 2010. « Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development. » Document de travail consacré à la recherche sur les politiques 5245, Banque mondiale, Washington, DC.

Behrman, Jere R. et John Hoddinott. 2001. « An Evaluation of the Impact of PROGRESA on Pre-school Child Height. » FCND Briefs 104, International Food Policy Research Institute, Washington, DC.

Belli, Pedro, Jock Anderson, Howard Barnum, John Dixon et Jee-Peng Tan. 2001. *Handbook of Economic Analysis of Investment Operations*. Washington DC : Banque mondiale.

Boardman, Anthony, Aidan Vining, David Greenberg et David Weimer. 2001. *Cost-Benefit Analysis: Concepts and Practice*. New Jersey: Prentice Hall.

Brent, Robert. 1996. *Applied Cost-Benefit Analysis*. Angleterre : Edward Elgar.

Fiszbein, Ariel, et Norbert Schady. 2009. *Conditional Cash Transfer, Reducing Present and Future Poverty*. World Bank Policy Research Report. Banque mondiale, Washington, DC.

Gertler, Paul J. 2004. « Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. » *American Economic Review* 94 (2) : 336–41.

Gertler, Paul J. et John W. Molyneaux. 1994. « How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility. » *Demography* 31 (1): 33–63.

———. 2000. « The Impact of Targeted Family Planning Programs in Indonesia. » *Population and Development Review 26* : 61–85.

Imas, Linda G. M. et Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington DC : Banque mondiale.

Kremer, Michael et Edward Miguel. 2004. « Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. » *Econometrica* 72 (1) : 159–217.

Kremer, Michael, Sylvie Moulin et Robert Namunyu. 2003. « Decentralization: A Cautionary Tale. » Poverty Action Lab Paper 10, Massachusetts Institute of Technology, Cambridge, MA.

Levy, Santiago et Evelyne Rodríguez. 2005. *Sin Herencia de Pobreza: El Programa Progresa-Oportunidades de México*. Washington DC : Banque interaméricaine de développement.

McKay, Harrison, Arlene McKay, Leonardo Siniestra, Hernando Gomez et Pascuala Lloreda. 1978. « Improving Cognitive Ability in Chronically Deprived Children. » *Science* 200 (21) : 270–78.

Poverty Action Lab. 2005. « Primary Education for All. » *Fighting Poverty: What Works?* N°1 (automne) : n.p. http://www.povertyactionlab.org.

Rao, Vijayendra et Michael Woolcock. 2003. « Integrating Qualitative and Quantitative Approaches in Program Evaluation. » *In The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, éd. F. J. Bourguignon and L. Pereira da Silva, 165–90. New York: Oxford University Press.

Schultz, Paul. 2004. « School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. » *Journal of Development Economics* 74 (1) : 199–250.

Skoufias, Emmanuel et Bonnie McClafferty. 2001. « Is *Progresa* Working? Summary of the Results of an Evaluation by IFPRI. » Institut international de recherche sur les politiques alimentaires, Washington, DC.

Vermeersch, Christel et Michael Kremer. 2005. « School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. » Document de travail consacré à la recherche sur les politiques 3523, Banque mondiale, Washington, DC.

Zerbe, Richard et Dwight Dively. 1994. *Benefit Cost Analysis in Theory and Practice*. New York : Harper Collins Publishing.

CHAPITRE 2

Formulation des questions d'évaluation

Le présent chapitre s'attèle aux premières étapes de l'élaboration d'une évaluation. Ces étapes comprennent la définition du type de questions auxquelles l'évaluation répond, la construction d'une théorie du changement montrant comment le projet est censé atteindre les résultats espérés, la réalisation d'une chaîne de résultats, la formulation des hypothèses qui seront testées par l'évaluation et la sélection d'indicateurs de performance.

Toutes ces étapes contribuent à formuler une question d'évaluation. Il est primordial qu'elles soient considérées dès la conception du programme, en étroite collaboration avec les parties prenantes, y compris les décideurs et responsables du programme, dans l'optique d'obtenir une vision commune des objectifs et de la manière de les atteindre. Un tel dialogue permet de forger un consensus sur les principales questions auxquelles l'évaluation répondra et de renforcer les liens entre l'évaluation, la mise en œuvre du programme et l'élaboration des politiques publiques. Ces étapes sont aussi essentielles pour établir la transparence et la spécificité nécessaires à la réalisation d'une bonne évaluation d'impact, tout comme elles contribuent en parallèle à la conception et à l'exécution d'un programme efficace. Chaque étape, de la formulation précise d'objectifs et de questions aux résultats espérés en passant par la formulation de la théorie du changement, est définie dans ce chapitre et articulée au sein d'une forme de modèle logique, la chaîne de résultats.

Types de questions d'évaluation

Toute évaluation débute par la formulation d'une question de recherche propre à la politique à l'étude. Le travail d'évaluation consiste ensuite à générer des arguments crédibles pour répondre à cette question. Comme nous l'expliquerons plus tard, la question fondamentale d'une évaluation d'impact peut être formulée ainsi : *quel est l'impact (ou effet causal) d'un programme sur un résultat donné ?* Pour l'un des exemples de la partie 2 de ce livre, la question d'étude s'articule comme suit : *quel est l'impact d'un programme de subvention de l'assurance maladie sur les dépenses de santé des ménages ?* La question peut également porter sur l'évaluation de plusieurs options de conception des programmes, par exemple : *quelle combinaison de campagnes de publipostage et de séances de conseils aux familles donne les meilleurs résultats lorsqu'il s'agit d'encourager l'allaitement maternel ?* La formulation d'une question d'évaluation claire et pertinente constitue le point de départ de toute évaluation efficace.

Théories du changement

Une théorie du changement est une description de la manière dont une intervention est censée produire les résultats espérés. Elle décrit la logique causale expliquant comment et pourquoi un projet, un programme ou une politique atteindra les résultats visés. L'existence d'une théorie du changement est fondamentale pour les évaluations d'impact étant donné l'importance qu'elles portent aux relations de cause à effet. La théorie du changement est l'une des premières étapes de la conception d'une évaluation, car elle contribue à la formulation des questions de recherche.

Les théories du changement décrivent une série d'évènements conduisant à un résultat ; elles énoncent les conditions et les hypothèses nécessaires pour que des changements se produisent ; elles mettent en évidence la logique causale sous-jacente au programme et inscrivent les interventions dans cette logique causale. Un travail conjoint entre les diverses parties prenantes pour définir une théorie du changement est souvent utile pour clarifier et améliorer l'élaboration du programme. Ceci est particulièrement important dans le cas des programmes qui visent à modifier des comportements : les théories du changement peuvent aider à décomposer les intrants et les activités constituant les interventions, les extrants qu'elles produisent et les résultats qui découlent des changements de comportement espérés des bénéficiaires.

Le début du processus de conception du programme constitue le meilleur moment pour formuler une théorie du changement ; les parties prenantes peuvent alors se réunir pour élaborer une vision commune du programme, de ses objectifs et des moyens à mettre en œuvre pour les atteindre. Les responsables peuvent ensuite implémenter le programme sur la base d'une compréhension commune de son fonctionnement et de ses objectifs.

Par ailleurs, il est important que les concepteurs du programme passent en revue la littérature au sujet d'interventions ou expériences similaires, et qu'ils vérifient soigneusement le contexte et les hypothèses qui sous-tendent la logique causale de la théorie du changement adoptée. Par exemple, pour le projet des sols en ciment au Mexique (voir encadré 2.1), la littérature existante permet de comprendre les mécanismes de transmission des parasites et la manière dont ils provoquent des diarrhées chez les enfants.

Encadré 2.1 : Théorie du changement

Des sols en ciment font le bonheur des Mexicains

Dans le cadre de leur évaluation du Projet Piso Firme ou « sol en dur », Cattaneo et al. (2009) étudient l'impact d'une amélioration de l'habitat sur la santé et le bien-être. Le projet, tout comme l'évaluation, repose sur une théorie du changement très claire.

L'objectif du Projet Piso Firme est d'améliorer le niveau de vie, notamment l'état de santé, de groupes vulnérables vivant dans des zones pauvres à forte densité de population. Le programme a d'abord été lancé dans le nord du pays, dans l'État de Coahuila, sur la base d'une appréciation du gouverneur Enrique Martínez et son équipe de campagne.

La chaîne de résultats du programme est claire. Une enquête porte-à-porte est effectuée auprès des ménages éligibles, et les ménages reçoivent l'équivalent de 50 m^2 de ciment. Les autorités assurent l'achat et la livraison du ciment tandis que les ménages et les volontaires des communautés fournissent la main-d'œuvre. L'extrant du programme est la construction, en une journée environ, d'un sol en ciment. Les résultats espérés de cette intervention sont, notamment, une amélioration de l'hygiène, de la santé et du bien-être des bénéficiaires.

La logique sous-jacente à cette chaîne de résultats est que les sols en terre battue sont des vecteurs de transmission des parasites, car ils sont plus difficiles à maintenir propres. Les parasites vivent et se reproduisent dans les excréments ; sont introduits dans les logements par les animaux, les enfants ou les chaussures, et peuvent être ingérés. Les données montrent que les enfants en bas âge vivant dans des maisons au sol en terre battue ont plus de risques d'être contaminés par des parasites intestinaux qui peuvent entraîner diarrhées et malnutrition, elles-mêmes responsables de retards dans le développement cognitif ou de décès. Les sols en ciment permettent d'interrompre la transmission parasitaire. Ils permettent en outre de mieux contrôler la température et sont plus esthétiques.

Ces résultats espérés contribuent à formuler les questions de recherche pour l'évaluation effectuée par Cattaneo et ses collaborateurs. Ils testent l'hypothèse que le remplacement des sols en terre battue par des sols en dur réduit l'incidence des diarrhées, de la malnutrition et des déficiences en oligo-éléments. Ils considèrent ensuite si ces changements entrainent aussi une amélioration du développement cognitif des enfants en bas âge. Les chercheurs examinent aussi si l'intervention améliore le bien-être des adultes tel que mesuré par le degré de satisfaction des personnes à l'égard de leur habitat et par la baisse du taux de dépression et de stress.

Source : Cattaneo et al. 2009.

Chaîne de résultats

Une théorie du changement peut être formalisée de différentes manières, par exemple par des modèles théoriques, des modèles logiques ou de chaînes de résultats[1]. Tous ces modèles comprennent les éléments fondamentaux d'une théorie du changement. En d'autres termes, ils articulent tous une chaîne causale, des conditions et des influences extérieures, et des hypothèses de base. Dans le présent ouvrage, nous allons nous concentrer sur la chaîne de résultats. Elle constitue, selon nous, le modèle le plus simple et le plus clair pour élaborer une théorie du changement dans le contexte opérationnel des programmes de développement.

Concept clé :
La chaîne de résultats établit la séquence d'intrants, d'activités et d'extrants contribuant à la réalisation des résultats intermédiaires et finaux espérés.

La chaîne de résultats est une représentation logique et plausible de la manière dont une séquence d'intrants, d'activités et d'extrants produits par un projet entre en interaction avec le comportement des bénéficiaires pour réaliser un impact donné (figure 2.1). Cette chaîne établit une logique causale du début à la fin du projet, depuis la mise à disposition des ressources jusqu'aux objectifs à long terme. Une chaîne de résultats est généralement composée des éléments suivants :

Intrants : ressources dont dispose le projet, y compris le personnel et le budget

Activités : actions entreprises ou travaux réalisés pour transformer les intrants en extrants

Extrants : biens et services tangibles produits par les activités du projet (les extrants sont sous le contrôle direct de l'agence chargée de l'exécution du programme)

Résultats intermédiaires : résultats susceptibles d'être atteints lorsque la population bénéficiaire utilise les extrants du projet (résultats généralement atteints à court et moyen terme)

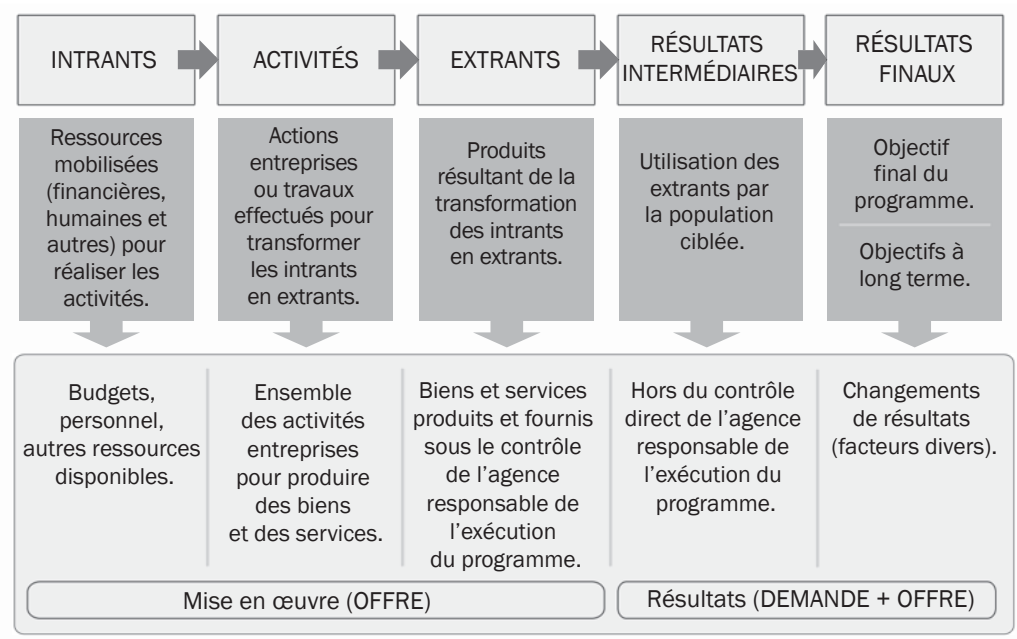
Résultats finaux : objectifs finaux du projet (ils peuvent subir l'influence de nombreux facteurs et sont généralement atteints à plus long terme).

Une chaîne de résultats comprend trois parties principales :

La mise en œuvre : travaux prévus réalisés par le projet, comprenant les intrants, les activités et les extrants. Il s'agit d'éléments dont l'agence responsable de l'exécution du programme peut faire un suivi direct dans le but de mesurer la performance du projet.

Les résultats : les résultats espérés comprennent les résultats intermédiaires et les résultats finaux. Ces résultats ne sont pas entièrement sous le contrôle direct de l'agence responsable de l'exécution du programme et sont tributaires des changements de comportement des bénéficiaires du programme. Autrement dit, ils dépendent de l'interaction entre l'offre (mise en œuvre) et la demande (bénéficiaires). Ce sont ces résultats qui font l'objet d'une évaluation d'impact en vue de mesurer l'efficacité du programme.

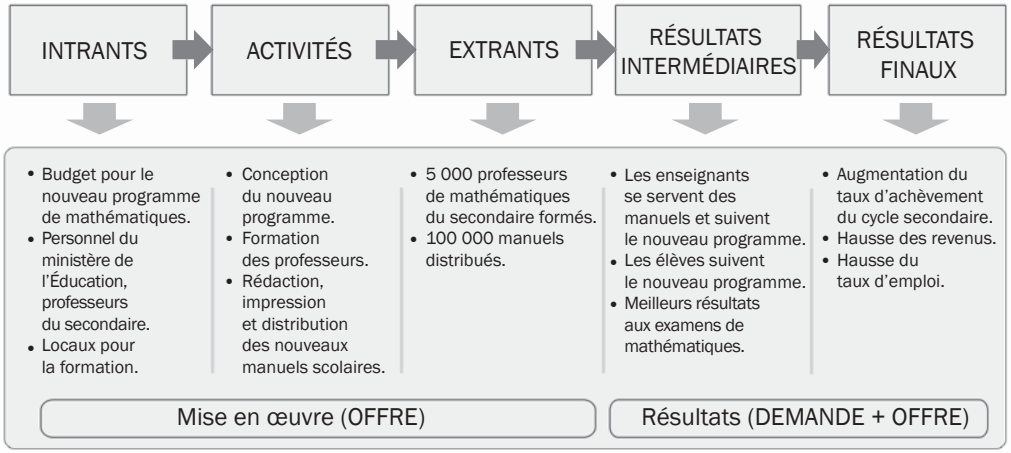
Figure 2.1 Qu'est ce qu'une chaîne de résultats ?



Les hypothèses et les risques : Les hypothèses et les risques ne sont pas présentés dans la figure 2.1. Ils comprennent toute information extraite de la littérature existante qui est pertinente pour la théorie du changement proposée, ainsi que les hypothèses sur lesquelles elle repose, des références aux résultats de programmes similaires, les risques qui pourraient remettre en cause les résultats espérés et toute stratégie mise en place pour atténuer ces risques.

Prenons l'exemple du ministère de l'Éducation d'un pays A qui souhaite lancer une nouvelle méthode d'enseignement des mathématiques dans le cycle secondaire. Comme l'illustre la figure 2.2, les intrants du programme se constituent du personnel du ministère, des enseignants du secondaire, des moyens financiers alloués au nouveau programme de mathématiques et des locaux pour organiser la formation des professeurs. Les activités comprennent la conception du nouveau programme de mathématiques, la préparation d'une formation pour les professeurs, la formation des professeurs ainsi que la commande, l'impression et la distribution des nouveaux manuels. Les extrants sont le nombre de professeurs formés, le nombre de manuels distribués dans les classes et l'adaptation des exa-

Figure 2.2 Chaîne de résultats d'un programme de mathématiques du cycle secondaire



mens de mathématiques au nouveau programme. Les résultats à court terme comprennent l'utilisation des nouvelles méthodes et des nouveaux manuels par les enseignants et l'adoption du nouveau curriculum. Les résultats à moyen terme sont l'amélioration des résultats des étudiants aux examens de mathématiques. Les résultats finaux incluent un taux accru d'étudiants terminant le cycle secondaire, une hausse du taux d'emploi et des revenus plus élevés des diplômés.

Les chaînes de résultats sont utiles pour tous les projets, qu'une évaluation d'impact soit prévue ou non. En effet, elles permettent aux décideurs et aux responsables de rendre explicites les objectifs du programme, de comprendre sa logique causale et de déterminer la séquence d'évènements sur laquelle repose son succès. De plus, les chaînes de résultats facilitent les discussions relatives au suivi et à l'évaluation, car elles mettent en exergue les informations qui doivent faire l'objet d'un suivi et les changements de résultats sur lesquels l'évaluation devra se concentrer.

Pour comparer différentes options de mise en œuvre d'un même programme, les chaînes de résultats peuvent être représentées sous la forme d'arbres de résultats indiquant toutes les alternatives envisagées au moment de la conception ou de la restructuration du programme. Ces arbres de résultats indiquent les différentes options stratégiques et opérationnelles qui peuvent mener aux objectifs spécifiques du programme ; ils peuvent servir de support de réflexion sur les options à tester et à évaluer. Par exemple, plusieurs interventions peuvent permettre de remplir l'objectif d'améliorer les connaissances dans le domaine financier, par exemple une campagne d'information ou une formation pour adultes.

Hypothèses pour l'évaluation

Après avoir constitué la chaîne de résultats, vous pouvez vous atteler à la formulation des hypothèses à tester dans le cadre de l'évaluation d'impact. Dans l'exemple du nouveau programme de mathématiques, les hypothèses pourraient être les suivantes :

- Le nouveau programme est supérieur à l'ancien pour améliorer les connaissances en mathématiques.
- Les enseignants formés utilisent le nouveau programme plus efficacement que les autres enseignants.
- Si la formation des enseignants et la distribution des manuels sont réalisées, les professeurs utiliseront ces manuels et adopteront le nouveau programme, et les étudiants suivront ce programme.
- Si la formation des enseignants et la distribution des manuels sont réalisées, les résultats aux examens de mathématiques augmenteront de cinq points en moyenne.
- Les résultats obtenus en mathématiques dans le secondaire ont une influence sur le taux d'achèvement du cycle secondaire et sur l'insertion professionnelle des étudiants.

Sélection des indicateurs de performance

Une chaîne de résultats clairement articulée est utile pour identifier les indicateurs à mesurer pour suivre et évaluer la performance des programmes. Ces indicateurs portent aussi bien sur le suivi de la mise en œuvre du programme que sur l'évaluation des résultats. Là encore, il est utile d'associer l'ensemble des parties prenantes au programme à la sélection des indicateurs afin qu'elles fournissent une mesure adéquate de sa performance. En règle générale, les indicateurs doivent être :

Concept clé :
Un bon indicateur est spécifique, mesurable, attribuable, réaliste et ciblé.

- *Spécifiques :* pour mesurer l'information nécessaire le plus précisément possible
- *Mesurables :* pour assurer que l'information puisse effectivement être obtenue
- *Attribuables :* pour pouvoir plausiblement attribuer chaque mesure effectuée aux efforts fournis dans le cadre du projet
- *Réalistes :* pour que les données puissent être obtenues à temps, à une fréquence et à un coût raisonnables
- *Ciblés :* pour que les indicateurs visent bien la population cible.

Il est important de définir des indicateurs tout au long de la chaîne de résultats sans se limiter aux résultats, de manière à pouvoir faire le suivi de toute la logique causale du programme. Même dans le cadre d'une évaluation d'impact, il est essentiel d'examiner les indicateurs de mise en œuvre des interventions pour s'assurer qu'elles ont été menées comme prévu, qu'elles ont touché les bénéficiaires visés et qu'elles ont été réalisées au moment opportun (voir Kusek et Rist, 2004, ou Imas et Rist, 2009 pour plus d'informations sur la sélection des indicateurs de performance). Faute d'indicateurs couvrant toute la chaîne des résultats, l'évaluation d'impact risque de devenir une « boîte noire » qui se limite à indiquer si les résultats attendus se sont matérialisés ou pas sans pour autant pouvoir expliquer pourquoi.

Outre la sélection des indicateurs, il est également important de définir d'où proviennent les données requises à la mesure des indicateurs de performance. Le tableau 2.1 récapitule les éléments de base d'un plan de suivi et évaluation ainsi que les modalités à suivre pour générer chacun des indicateurs de manière fiable et opportune.

Tableau 2.1 Éléments d'un plan de suivi et évaluation

Élément	Description
Résultats espérés (résultats et extrants)	Obtenus à partir des documents de conception du programme et de la chaîne de résultats.
Indicateurs (avec valeurs dans les données de base et objectifs indicatifs)	Tirés de la chaîne des résultats ; les indicateurs doivent être Spécifiques, Mesurables, Attribuables, Réalistes, Ciblés.
Source des données	Sources ou lieu où les données seront recueillies, par exemple un rapport, ou une réunion des parties prenantes au projet.
Fréquence des données	Fréquence de disponibilité des données.
Responsabilités	Qui est responsable de l'organisation de la collecte des données ainsi que de la vérification de la qualité des données et des sources ?
Analyse et compte rendu	Fréquence des analyses, méthode d'analyse et responsabilité du compte rendu.
Ressources	Estimation des ressources nécessaires et engagées pour réaliser les activités de suivi et évaluation.
Utilisation finale	Qui recevra les informations et les utilisera ? Dans quel but ?
Risques	Quels sont les hypothèses et les risques liés aux activités de suivi et d'évaluation ? Comment peuvent-ils affecter les activités de suivi et évaluation prévues ainsi que la qualité des données ?

Source : adapté d'une publication du PNUD, 2009.

Feuille de route pour les parties 2 et 3

Dans cette première partie de l'ouvrage, nous avons exposé pourquoi réaliser des évaluations d'impact et quand les mettre en œuvre. Nous avons évoqué les divers objectifs des évaluations d'impact ainsi que les questions fondamentales de politique auxquelles elles répondent. Nous avons souligné la nécessité de bien définir la théorie du changement pour indiquer les mécanismes par lesquels un programme a un impact sur les résultats finaux. Le but de l'évaluation d'impact est essentiellement de vérifier si cette théorie du changement s'applique ou non dans les faits.

La partie 2, intitulée *Comment évaluer ?*, porte sur les diverses méthodes qui permettent de constituer des groupes de comparaison adéquats et réaliser une évaluation valide des impacts d'un programme. Nous commençons par introduire le *contrefactuel*, notion fondamentale à toute évaluation d'impact, en mettant l'accent sur les propriétés de l'estimation du contrefactuel et en donnant des exemples de contrefactuels non valides ou contrefaits. Nous présentons ensuite diverses méthodes pour obtenir une estimation valable du contrefactuel. Nous évoquons notamment l'intuition sous-jacente de quatre catégories de méthodologies : *la sélection aléatoire, le modèle de discontinuité de la régression, la double différence et l'appariement*. Nous étudions les circonstances dans lesquelles chaque méthode fournit une estimation valable du contrefactuel, le contexte opérationnel dans lequel ces méthodes sont appropriées et leurs principales limites. Tout au long de la deuxième partie du manuel, une étude de cas (le Programme de subvention de l'assurance maladie) est utilisée pour illustrer les diverses méthodes. Nous présentons aussi des exemples concrets de l'application de chacune des méthodes à des programmes de développement.

La partie 3 est consacrée aux étapes à suivre pour mettre en œuvre, gérer ou commissionner une évaluation d'impact. À ce stade, nous considérons que les objectifs de l'évaluation ont été définis, qu'une théorie du changement a été élaborée et que les questions d'évaluation ont été formulées. Nous passons en revue les principaux points à résoudre au moment d'élaborer le plan de l'évaluation d'impact. Nous présentons des règles simples pour choisir le groupe de comparaison le plus approprié dans un contexte donné. Nous établissons un cadre qui aide à choisir, parmi les méthodes d'évaluation présentées dans la partie 2, la méthode la mieux adaptée à un programme en fonction de ses règles opérationnelles. Nous passons ensuite en revue les quatre grandes phases de la réalisation d'une évaluation : mettre en œuvre l'évaluation, choisir un échantillon, collecter les données, produire et diffuser les conclusions.

Note

1. University of Wisconsin-Extension (2010) propose des informations détaillées sur la manière d'articuler une chaîne de résultats, ainsi qu'une liste complète de références. Imas et Rist (2009) présentent une revue plus complète des théories du changement.

Références

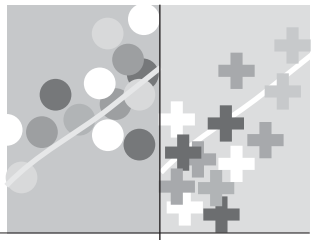
Cattaneo, Matias, Sebastian Galiani, Paul Gertler, Sebastian Martinez et Rocio Titiunik. 2009. « Housing, Health and Happiness. » *American Economic Journal : Economic Policy* 1 (1) : 75–105.

Imas, Linda G. M. et Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington DC : Banque mondiale.

Kusek, Jody Zall et Ray C. Rist. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington DC : Banque mondiale.

PNUD (Programme des Nations Unies pour le développement). 2009. *Guide de la planification, du suivi et de l'évaluation axés sur les résultats du développement*. New York : PNUD.

University of Wisconsin-Extension. 2010. « Enhancing Program Performance with Logic Models. » Cours en ligne. http://www.uwex.edu/ces/pdande/evaluation/evallogicmodel.html.



Partie 2

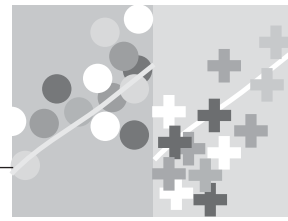
COMMENT ÉVALUER ?

Maintenant que nous avons souligné les raisons d'évaluer l'impact des programmes et des politiques publiques, cette deuxième partie examine comment procèdent les évaluations d'impact, les questions auxquelles elles répondent, les méthodes d'évaluation à disposition ainsi que les avantages et les inconvénients de chacune d'elles. Nous nous intéressons notamment aux méthodes de sélection aléatoire, au modèle de discontinuité de la régression, à la double différence et à l'appariement.

Comme l'expose la première partie, une évaluation d'impact vise à établir et à quantifier l'impact d'une intervention sur les résultats qui intéressent les analystes et les décideurs politiques. Dans cette deuxième partie du manuel, nous présentons une étude de cas : « le programme de subvention de l'assurance maladie » (PSAM). Nous répondons à plusieurs reprises à la même question concernant l'évaluation d'impact du PSAM à partir des mêmes sources de données, mais en utilisant différentes méthodes qui fournissent des réponses différentes, parfois même opposées. (Nous supposons ici que les données ont

été totalement dépurées). Votre tâche est d'identifier les raisons pour lesquelles les estimations d'impact du PSAM diffèrent selon la méthode d'évaluation retenue et de déterminer les résultats que vous estimez suffisamment fiables pour fournir des recommandations stratégiques de politiques publiques.

Le contexte de l'étude de cas du PSAM est le suivant : les autorités entament un programme de réformes du secteur de la santé de grande envergure dans le but d'améliorer l'état de santé de la population. L'objectif général de ces réformes est d'améliorer l'accès aux services de santé et leur qualité dans les régions rurales pour atteindre un niveau similaire aux zones urbaines. Le PSAM est un projet pilote novateur potentiellement fort coûteux. Le programme subventionne le système d'assurance maladie pour qu'il couvre le coût des soins de santé primaires et des médicaments pour les ménages ruraux pauvres. L'objectif principal du PSAM est de réduire le coût des soins de santé pour les ménages pauvres et, en définitive, d'améliorer les résultats en matière de santé. Les autorités envisagent d'étendre le PSAM à l'ensemble du pays. Cette décision coûterait des centaines de millions de dollars, mais les décideurs craignent que sans subvention, les ménages ruraux pauvres ne soient pas en mesure de payer les soins de santé de base, ce qui aurait des conséquences néfastes sur leur état de santé. Dans ce contexte, la question clé d'évaluation est la suivante : *quel est l'impact du PSAM sur les dépenses en soins de santé à la charge des ménages et sur l'état de santé des familles pauvres ?* La réponse à de telles questions permet d'orienter les décideurs dans leurs choix de politiques à adopter et de programmes à mettre en œuvre. À leur tour, ces programmes peuvent avoir un impact sur le bien-être de millions de personnes dans le monde. Les questions d'évaluation d'impact sont donc particulièrement importantes, et cette partie du manuel passe en revue comment y répondre de manière rigoureuse.



CHAPITRE 3

Inférence causale et contrefactuel

Nous allons tout d'abord examiner deux concepts essentiels pour réaliser des évaluations précises et fiables, à savoir l'inférence causale et le contrefactuel.

Inférence causale

La question fondamentale de l'évaluation d'impact constitue essentiellement un problème d'*inférence causale*. Évaluer l'impact d'un programme sur une série de résultats revient à évaluer l'effet causal du programme sur lesdits résultats. La plupart des questions de politique invoquent des relations de cause à effet : la formation des professeurs *entraîne*-t-elle une amélioration des résultats des élèves aux examens ? Les programmes de transferts monétaires conditionnels *entraînent*-ils une amélioration de l'état de santé des enfants ? Les programmes de formation professionnelle *entraînent*-ils une amélioration des revenus des bénéficiaires ?

Même si les questions qui abordent une relation de cause à effet sont courantes, il n'est jamais facile d'établir qu'une relation est effectivement causale. Par exemple, le simple fait d'observer que le revenu des bénéficiaires d'un programme de formation professionnelle augmente ne suffit pas à établir un lien de causalité. Le revenu d'un bénéficiaire pourrait en effet avoir augmenté même s'il n'avait pas suivi le programme de formation grâce, par exemple, à ses propres efforts, à l'évolution des conditions sur le marché du travail ou à tout autre facteur susceptible d'avoir un impact sur le revenu à travers le temps. Les évaluations d'impact permettent d'établir un lien de causalité en démontrant empiriquement dans quelle mesure un pro-

gramme donné — *et uniquement ce programme* — a contribué à changer un résultat. Pour établir un lien de causalité entre un programme et un résultat, nous utilisons des méthodes d'évaluation d'impact qui permettent d'écarter la possibilité que des facteurs autres que le programme à l'étude puissent expliquer l'impact observé.

La réponse à la question fondamentale de l'évaluation d'impact, à savoir *quel est l'impact ou l'effet causal d'un programme P sur un résultat Y*, est donnée par la formule de base d'évaluation d'impact :

$$\alpha = (Y \mid P = 1) - (Y \mid P = 0).$$

Selon cette formule, l'effet causal α d'un programme (*P*) sur un résultat (*Y*) est la différence entre le résultat (Y) obtenu avec le programme (autrement dit avec $P = 1$) et le même résultat (*Y*) obtenu sans le programme (c.-à-d. avec $P = 0$).

Par exemple, si *P* est un programme de formation professionnelle et *Y* le revenu, l'effet causal du programme de formation professionnelle α est la différence entre le revenu d'une personne donnée (*Y*) après avoir participé au programme de formation (donc avec $P = 1$) et le revenu qu'aurait eu la même personne (*Y*) au même moment si elle n'avait pas participé au programme (avec $P = 0$). Autrement dit, nous cherchons à mesurer le revenu au même moment et pour la même unité d'observation (une personne dans le cas présent), mais dans deux cas de figure différents. S'il était possible de procéder ainsi, nous pourrions observer le revenu gagné par une même personne au même moment à la fois après avoir suivi le programme de formation professionnelle et sans l'avoir suivi, de manière à ce que toute différence de revenu pour cette personne ne puisse s'expliquer que par sa participation au programme. En comparant une même personne à elle-même au même moment avec et sans le programme, nous serions capables d'éliminer tout facteur externe susceptible de contribuer à la différence de revenu. Nous pourrions alors conclure sans aucun doute que la relation entre le programme de formation professionnelle et le revenu est bel et bien causale.

La formule de base d'évaluation d'impact est valable pour toute unité à l'étude, qu'il s'agisse d'une personne, d'un ménage, d'une communauté, d'une entreprise, d'une école, d'un hôpital ou de toute autre unité d'observation qui peut bénéficier d'un programme. Cette formule est également applicable à tout indicateur de résultat (*Y*) qu'un programme en place peut de manière plausible affecter. Si nous parvenons à mesurer les deux éléments clés de cette formule, à savoir le résultat (*Y*) à la fois en présence et en l'absence du programme, nous pourrons alors répondre à n'importe quelle question sur l'impact de ce programme.

Contrefactuel

Comme nous l'avons vu ci-dessus, l'impact α d'un programme est conceptuellement la différence du résultat (*Y*) pour une même personne lorsqu'elle bénéficie d'un programme (*P*) et n'en bénéficie pas. Pourtant, il est bien évidemment impossible d'observer la même personne au même moment dans deux cas de figure

différents. Une personne ne peut pas simultanément participer à un programme et ne pas y participer. La personne ne peut donc pas être observée au même moment dans les deux cas de figure (autrement dit, en tant que bénéficiaire et non-bénéficiaire du programme). Ce problème s'appelle le « problème contrefactuel » : comment mesurer ce qui se serait passé dans d'autres circonstances ? Nous pouvons certes observer et mesurer le résultat (Y) pour les participants au programme ($Y \mid P = 1$), mais il n'existe aucune donnée pour déterminer ce qu'auraient été les résultats pour un bénéficiaire en l'absence du programme ($Y \mid P = 0$). Dans la formule de base d'évaluation d'impact, le terme ($Y \mid P = 0$) *représente le contrefactuel*. Le contrefactuel peut être considéré comme *ce qui serait arrivé* si un participant n'avait en réalité pas bénéficié du programme. Autrement dit, le contrefactuel est le résultat (Y) qui aurait été obtenu en l'absence de programme (P).

Concept clé :
Le contrefactuel est une estimation de ce qu'aurait été le résultat (Y) pour un bénéficiaire du programme en l'absence du programme (P).

Prenons l'exemple de « Monsieur Malchance » qui avale un comprimé rouge et décède cinq jours plus tard. Nous ne pouvons pas conclure que le comprimé rouge *a causé* la mort de M. Malchance uniquement parce que celui-ci est décédé après avoir pris un comprimé. M. Malchance était peut-être très malade lorsqu'il a avalé ce comprimé rouge, auquel cas il est possible que sa maladie et non le comprimé ait provoqué son décès. Pour inférer un lien de causalité, il faudra écarter tout autre facteur susceptible d'avoir tenu un rôle dans le résultat, en l'occurrence le décès de M. Malchance. Dans cet exemple, il s'agira de déterminer ce qui se serait passé si M. Malchance *n'avait pas* pris ce comprimé. Toutefois, étant donné que M. Malchance a effectivement pris le comprimé rouge, il n'est pas possible d'observer directement ce qui serait arrivé s'il ne l'avait pas fait. Ce qui lui serait arrivé s'il n'avait pas pris le comprimé rouge constitue le contrefactuel. Le principal défi pour un évaluateur est justement de déterminer à quoi ressemble un contrefactuel (voir l'encadré 3.1).

Dans le cadre d'une évaluation d'impact, il est relativement facile de mesurer le premier terme de la formule de base ($Y \mid P = 1$), c'est-à-dire le résultat du groupe recevant le traitement. Il suffit de mesurer le résultat pour la population ayant bénéficié du programme. En revanche, le second terme de l'équation ($Y \mid P = 0$) ne peut pas être observé directement auprès des bénéficiaires du programme ; il faut donc reconstituer les éléments manquants en *estimant le contrefactuel*. Pour ce faire, nous avons recours à des *groupes de comparaison* (ou « groupes témoins »). Le reste de la partie 2 du manuel est consacré aux différentes méthodes ou approches qui peuvent être utilisées pour concevoir des groupes de comparaison valides, reproduisant ou imitant avec précision le contrefactuel. L'identification de ces groupes de comparaison est la pierre angulaire de toute évaluation d'impact, quel que soit le type de programme à évaluer. Autrement dit, sans contrefactuel valide, l'impact d'un programme ne peut pas être établi.

Encadré 3.1 : Estimation du contrefactuel

Mademoiselle Unique et le programme de transferts monétaires conditionnels

Mademoiselle Unique est un bébé dont la maman se voit offrir un transfert monétaire mensuel à condition qu'elle assure que sa petite Unique soit vaccinée, effectue régulièrement un bilan de santé et un suivi de la croissance au centre de santé local. Les autorités pensent que le transfert monétaire incitera la maman de Mademoiselle Unique à recourir aux services de santé, condition préalable pour bénéficier du programme, et que cela permettra à Mademoiselle Unique de devenir une grande fille en bonne santé. Pour l'évaluation d'impact, les autorités choisissent la taille comme un indicateur de la santé à long terme. Supposons que Mademoiselle Unique soit mesurée à l'âge de trois ans. Si vous voulez évaluer l'impact du programme, l'idéal serait de pouvoir mesurer Mademoiselle Unique à l'âge de trois ans dans le cas de figure où sa mère bénéficie du transfert monétaire, et de mesurer la même demoiselle toujours à l'âge de trois ans, mais cette fois dans le cas de figure où sa maman ne reçoit aucune allocation. Vous pourriez alors comparer les deux tailles. S'il était possible de comparer la taille de Mademoiselle Unique à l'âge de trois ans alors que sa maman bénéficie du programme et sa taille au même âge en l'absence du programme, vous seriez certain que toute différence de taille serait due uniquement à la mise en place du programme. Toutes choses étant égales par ailleurs, pour Mademoiselle Unique, aucun autre facteur ne pourrait expliquer une éventuelle différence de taille dans les deux cas de figure.

Malheureusement, il est impossible d'observer la taille de Mademoiselle Unique à la fois en présence et en l'absence du programme de transferts monétaires : en effet, soit sa famille bénéficie du programme, soit elle n'en bénéficie pas. Autrement dit, on ne peut pas observer le contrefactuel. La maman de Mlle Unique ayant bénéficié du programme de transferts monétaires, nous ne pouvons pas savoir quelle aurait été la taille de sa fille en l'absence du programme. Or, trouver une comparaison pour Mademoiselle Unique constitue un véritable défi, cette demoiselle étant, bien sûr, unique. Son profil socioéconomique et ses caractéristiques génétiques et personnelles exactes ne peuvent se retrouver en aucune autre personne. Si nous comparions Mademoiselle Unique à un enfant (par exemple M. Inimitable) qui ne bénéficie pas du programme de transferts monétaires, la comparaison pourrait ne pas être valable. Mlle Unique n'est pas identique à M. Inimitable. Mlle Unique et M. Inimitable peuvent ne pas se ressembler, ils peuvent ne pas vivre au même endroit, ils peuvent ne pas avoir les mêmes parents et ils peuvent ne pas avoir eu la même taille à leur naissance. Donc, si nous observons que M. Inimitable est moins grand que Mlle Unique à l'âge de trois ans, nous ne pouvons pas savoir si cette différence de taille est due au programme de transferts monétaires ou à l'une des nombreuses différences qui peuvent exister entre ces deux enfants.

Estimation du contrefactuel

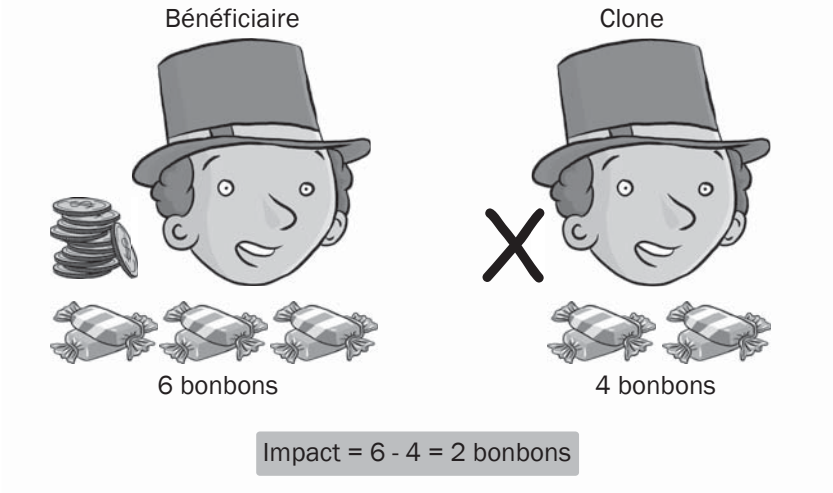
Pour illustrer l'estimation du contrefactuel, nous allons prendre un exemple qui, bien que sans importance sur le plan politique, nous permettra de mieux appréhender cette notion clé. Conceptuellement, pour résoudre le problème du contrefactuel,

l'évaluateur doit trouver le « clone parfait » pour chaque participant à un programme (figure 3.1). Par exemple, supposons que Fulanito perçoive 12 dollars supplémentaires d'argent de poche. Nous voudrions mesurer l'impact de cette augmentation d'argent de poche sur sa consommation de bonbons. S'il existait un clone parfait de Fulanito, l'évaluation serait aisée : il suffirait de comparer le nombre de bonbons consommés par Fulanito (disons six) avec le nombre de bonbons consommés par son clone ne recevant pas d'argent de poche supplémentaire (disons quatre). Dans ce cas, l'impact de l'argent de poche serait la différence entre ces deux chiffres (deux bonbons). Dans la réalité, les clones parfaits n'existent évidemment pas. Des différences importantes existent même entre les vrais jumeaux ayant un patrimoine génétique semblable.

Toutefois, même s'il est impossible de trouver un clone parfait pour chacun des bénéficiaires d'un programme, certains outils statistiques permettent de générer deux groupes qui, s'ils sont composés d'un nombre assez important d'individus, sont statistiquement indiscernables l'un de l'autre. Dans la pratique, l'un des objectifs clés d'une évaluation d'impact est d'identifier un groupe de participants au programme (groupe de traitement) et un groupe de non participants (groupe de comparaison) statistiquement identiques en l'absence du programme. Si les deux groupes sont identiques à la seule exception que l'un des groupes participe au programme et l'autre non, toute différence entre les résultats des deux groupes est attribuable au programme.

Le principal défi est alors de trouver un groupe de comparaison valide ayant les mêmes caractéristiques que le groupe de traitement. Plus précisément, le groupe de traitement et le groupe de comparaison doivent être semblables en au moins trois points. En premier lieu, les groupes de traitement et de comparaison doivent être

Figure 3.1 Le clone parfait



identiques en l'absence du programme. Il n'est pas nécessaire que toutes les unités du groupe de traitement soient identiques à toutes celles du groupe de comparaison, mais en moyenne, les caractéristiques des deux groupes doivent être les mêmes. Par exemple, l'âge moyen dans le groupe de traitement doit être le même que l'âge moyen dans le groupe de comparaison. En deuxième lieu, les deux groupes doivent réagir de la même manière au programme. Par exemple, le revenu des unités du groupe de traitement doit potentiellement augmenter à la suite d'un programme de formation dans la même mesure que celui des unités du groupe de comparaison si celles-ci avaient aussi reçu le programme. En troisième lieu, les groupes de traitement et de comparaison ne doivent pas être exposés de manière différenciée à d'autres interventions au cours de la période d'évaluation. Par exemple, si nous voulons évaluer l'impact de l'octroi supplémentaire d'argent de poche sur la consommation de bonbons, le groupe de traitement ne doit pas avoir été invité à se rendre au magasin de bonbons plus de fois que le groupe de comparaison, car il deviendrait alors difficile de distinguer les effets de l'accès accru aux bonbons des effets de l'augmentation du montant d'argent de poche.

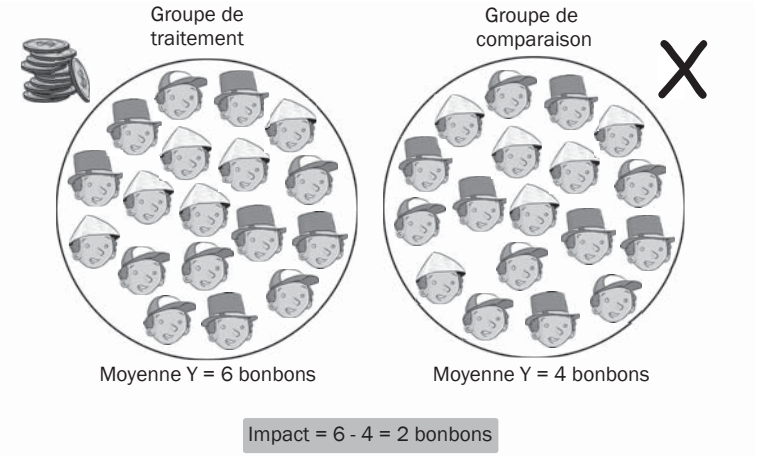
Concept clé :
Un groupe de comparaison valide doit avoir les mêmes caractéristiques que le groupe de participants au programme (« groupe de traitement ») à la seule différence que les unités du groupe de comparaison ne bénéficient pas du programme.

Quand ces trois conditions sont réunies, seul le programme peut expliquer les différences de résultat (Y) entre les deux groupes après sa mise en œuvre. Ceci est dû au fait que la seule différence entre le groupe de traitement et le groupe de comparaison est que les membres du groupe de traitement bénéficient du programme, mais pas les membres du groupe de comparaison. Quand les différences de résultat peuvent être totalement attribuées au programme, l'effet causal du programme est établi. Ainsi, au lieu de s'intéresser uniquement à l'impact de l'octroi supplémentaire d'argent de poche à Fulanito, il est possible d'analyser l'impact pour tout un groupe d'enfants (figure 3.2). Si vous pouvez identifier un autre groupe d'enfants totalement similaire, à la seule différence qu'ils ne recevront pas d'argent de poche supplémentaire, votre estimation de l'impact du programme sera alors la différence de consommation moyenne de bonbons entre les deux groupes. Par exemple, si la consommation moyenne du *groupe de traitement* est de six bonbons par enfant et celle du *groupe de comparaison* de quatre bonbons, l'impact moyen de l'octroi d'argent de poche supplémentaire sur la consommation de bonbons est de deux.

Concept clé :
Quand le groupe de comparaison n'est pas valide, l'estimation de l'impact du programme ne sera pas valide non plus : elle ne permettra pas d'estimer l'impact causal réel du programme. En termes statistiques, l'estimation est dite « biaisée ».

Maintenant que nous avons défini ce qu'est un *groupe de comparaison valide*, considérons les implications de mener une évaluation sans un tel groupe. Intuitivement, un groupe de comparaison non valide est un groupe qui diffère du groupe de traitement autrement que par la seule absence du traitement à l'étude. Ces autres différences peuvent rendre l'estimation d'impact invalide ou, en termes statistiques, *biaisée*. En effet, en présence d'autres différences entre les groupes de traitement et de comparaison, l'estimation ne permettra pas de déterminer l'impact réel du programme, car elle confondra l'effet du programme avec les effets des autres différences.

Figure 3.2 Un groupe de comparaison valide



Deux types d'estimation d'impact

Après avoir estimé l'impact du programme, l'évaluateur doit interpréter les résultats correctement. Une évaluation consiste toujours à estimer l'impact d'un programme en comparant les résultats obtenus par le groupe de traitement avec les estimations du contrefactuel obtenues d'un groupe de comparaison valide, comme indiqué par la formule de base d'évaluation d'impact. L'interprétation de l'impact du programme peut varier en fonction de ce que le traitement et le contrefactuel représentent réellement.

L'impact estimé α s'appelle l'estimation de l'« intention de traiter » (IDT) lorsque la formule de base est appliquée aux unités auxquelles le programme a été offert, qu'elles y participent effectivement ou non. L'Intention de traiter (IDT) est importante dans les cas où nous essayons de déterminer l'impact moyen d'un programme sur la population *ciblée* par le programme. Par contre, l'impact estimé α est appelé effet du « traitement sur les traités » (TT) lorsque la formule de base de l'évaluation d'impact est appliquée aux unités auxquelles le programme a été proposé et qui y ont effectivement participé. Les estimateurs IDT et TT seront identiques en cas d'adhérence totale, c'est-à-dire si toutes les unités auxquelles le programme a été proposé décident d'y participer. Nous reviendrons en détail sur la différence entre l'IDT et le TT mais nous pouvons d'ores et déjà commencer par un exemple.

Reprenons l'exemple du programme de subvention de l'assurance maladie (PSAM) évoqué en introduction de la partie 2 et au titre duquel chaque ménage du village bénéficiant du programme (village traité) peut s'inscrire pour recevoir un subside pour l'assurance maladie. Même si tous les ménages des villages traités sont

éligibles au programme, une partie d'entre eux (disons 10 %) peuvent décider de ne pas y participer (peut-être parce qu'ils ont déjà une assurance par le biais de leur travail, parce qu'ils sont en bonne santé et ne pensent pas qu'ils auront besoin de soins à l'avenir ou pour toute autre raison). Dans cet exemple, 90 % des ménages des villages traités décident de participer au programme et ont effectivement recours aux services du programme. Dans ce cas, l'estimateur IDT est obtenu en appliquant la formule de base d'évaluation d'impact à l'ensemble des ménages auxquels le programme a été proposé, autrement dit tous les ménages des villages traités. En revanche, l'estimation TT serait obtenue en appliquant la formule de base d'évaluation d'impact pour le sous-groupe des ménages qui décident de participer au programme, en l'occurrence 90 % des ménages traités.

Deux contrefactuels contrefaits

Dans la suite de la partie 2 du manuel, nous passerons en revue diverses méthodes qui peuvent être utilisées pour créer un groupe de comparaison valide afin d'estimer le contrefactuel. Auparavant, il est toutefois indispensable d'évoquer deux méthodes courantes, mais très risquées, de former des groupes de comparaison. Ces deux méthodes conduisent souvent à une estimation inappropriée du contrefactuel. Ces deux contrefactuels « contrefaits » sont 1) la comparaison *avant-après*, ou pré-post, qui compare les résultats pour le groupe de participants au programme avant et après la mise en œuvre du programme, et 2) la comparaison *avec-sans*, qui compare des unités ayant choisi de participer au programme avec des unités ayant choisi de ne pas y participer.

Contrefactuel contrefait 1 : comparaison avant-après

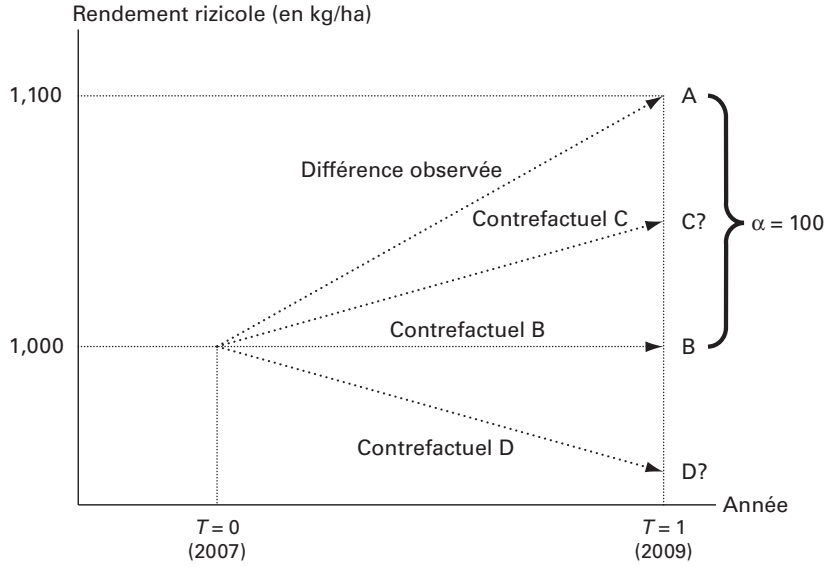
Une comparaison avant-après consiste à déterminer l'impact d'un programme en examinant l'évolution des résultats pour les participants au programme au fil du temps. Pour revenir à notre formule de base, le résultat pour le groupe de traitement ($Y \mid P = 1$) est alors tout simplement le résultat après l'intervention, alors que le contrefactuel ($Y \mid P = 0$) est estimé à partir du résultat avant l'intervention. Essentiellement, la comparaison repose sur l'hypothèse que si le programme n'avait pas existé, le résultat (Y) pour les participants au programme aurait été exactement le même qu'avant leur participation au programme. Malheureusement, dans la grande majorité des cas, cette hypothèse n'est pas valable.

Prenons l'exemple d'un programme de microfinance destiné aux agriculteurs pauvres en milieu rural. Ce programme propose des microcrédits aux agriculteurs pour leur permettre d'acheter des engrais afin d'accroître leur production de riz.

On sait que l'année précédant le lancement du programme, la production moyenne de riz était de 1 000 kg par hectare. Le programme de microfinance est lancé et l'année suivante les rendements passent à 1 100 kg par hectare. Si nous cherchons à mesurer l'impact du programme en nous fondant sur une comparaison avant-après, c'est le résultat avant intervention qui constituera le contrefactuel. En appliquant la formule de base, nous conclurions que le programme a permis une augmentation des rendements rizicoles de 100 kg par hectare.

Toutefois, imaginons que les précipitations étaient normales l'année précédant le lancement du programme, mais qu'une sécheresse a lieu l'année où le programme débute. Dans ce cas, nous ne pouvons pas considérer le résultat avant l'intervention comme un contrefactuel fiable. La figure 3.3 en décrit les raisons. Puisque les agriculteurs ont bénéficié du programme lors d'une année de la sécheresse, leur rendement moyen aurait été inférieur sans le programme de microfinance, par exemple au niveau D et non au niveau B comme le laisserait croire la comparaison avant-après. Dans ce cas, l'impact réel du programme est supérieur à 100 kg. À l'inverse, si les conditions climatiques avaient été meilleures, le rendement contrefactuel aurait pu être au niveau C. L'impact réel du programme aurait alors été inférieur à 100 kg.

Figure 3.3 Estimations avant et après d'un programme de microfinance



Autrement dit, à moins de pouvoir contrôler statistiquement pour le climat et *tous les autres facteurs* pouvant influencer les rendements rizicoles, il n'est pas possible de déterminer avec certitude l'impact réel du programme en faisant une comparaison avant-après.

Même si les comparaisons avant-après sont rarement valides pour réaliser une évaluation d'impact, elles restent utiles à d'autres fins. Les bases de données administratives de nombreux programmes enregistrent des informations sur les participants au fil du temps. Par exemple, un système de gestion de l'information dans le secteur éducatif peut collecter régulièrement des informations sur les taux de scolarisation dans les écoles où un programme de distribution de repas est en œuvre. Ces données permettent aux gestionnaires de programme de constater si le nombre d'enfants scolarisés augmente dans le temps. Ces informations sont importantes et tout à fait pertinentes pour les gestionnaires s'occupant de la planification et du suivi du secteur éducatif. Toutefois, conclure que le programme de distribution de repas dans les écoles est la *cause* du changement observé du taux de scolarisation serait risqué, car d'autres facteurs peuvent avoir affecté ce taux. Par conséquent, même s'il est très utile de suivre les changements d'indicateurs de résultat dans le temps pour un groupe de participants à un programme, il est généralement impossible de conclure en toute certitude que c'est effectivement le programme qui est la cause de l'amélioration observée (ni dans quelle mesure le programme y contribue) en présence d'autres facteurs variables dans le temps susceptibles d'avoir aussi influencé le résultat.

Comme nous l'avons vu avec l'exemple du programme de microfinance et des rendements rizicoles, les rendements peuvent être affectés par de nombreux facteurs variables dans le temps. De la même manière, une multitude de facteurs peuvent affecter les résultats que les programmes de développement visent à améliorer. Pour cette raison, le résultat avant la mise en œuvre d'un programme ne constitue pratiquement jamais une bonne estimation du contrefactuel. Nous qualifions donc la comparaison avant-après de « contrefactuel contrefait ».

Évaluation avant-après du Programme de subvention de l'assurance maladie (PSAM)

Pour mémoire, le PSAM est un nouveau programme de subvention de l'assurance maladie pour les ménages ruraux pauvres. Cette assurance couvre les dépenses relatives aux soins de santé primaires et à l'achat de médicaments. L'objectif du PSAM est de réduire le coût des soins de santé à la charge directe des ménages pauvres et, en définitive, d'améliorer les indicateurs de santé des bénéficiaires. De nombreux indicateurs de résultat peuvent être retenus pour évaluer l'impact du programme, mais en l'occurrence les autorités veulent surtout connaître les effets du PSAM sur les dépenses en soins primaires et en médicaments des familles pauvres, plus précisément sur les dépenses annuelles directes par personne (désignées ci-après par « dépenses de santé »).

Le PSAM représentera une part conséquente du budget national s'il est élargi à l'ensemble du pays (jusqu'à 1,5 % du PIB selon certaines estimations). De plus, la gestion d'un programme de cette nature est très complexe sur le plan administratif et logistique. Il a donc été décidé au plus haut niveau de l'État de lancer le PSAM tout d'abord sous la forme d'un programme pilote et de l'élargir progressivement en fonction des résultats obtenus lors de la première phase. À partir des résultats des analyses financières et coût-bénéfice, la présidente et les membres de son cabinet ont annoncé que pour être considéré comme viable et être étendu à tout le pays, le PSAM devait réduire les dépenses de santé annuelles moyennes par habitant d'au moins neuf dollars par rapport à ce qu'elles auraient été en l'absence du programme, et ce dans un délai de deux ans.

Le PSAM sera mis en œuvre dans 100 localités rurales au cours de la phase pilote. Juste avant le lancement du programme, votre gouvernement engage une société pour mener une enquête de référence des 4 959 ménages que comptent ces villages. L'enquête collecte des informations détaillées sur tous les ménages, y compris sur leur composition, leurs actifs, l'accès aux services de santé et les dépenses de santé au cours de l'année écoulée. Peu après la conduite de cette enquête de référence, le PSAM est lancé en fanfare dans 100 villages pilotes, accompagnés d'évènements communautaires et de campagnes promotionnelles pour encourager les ménages éligibles à participer.

Sur les 4 959 ménages de l'échantillon de référence, 2 907 s'inscrivent au PSAM au cours des deux premières années du programme. En deux ans, le PSAM donne de bons résultats selon plusieurs indicateurs. Les taux de couverture sont élevés et les enquêtes montrent que la plupart des ménages inscrits sont satisfaits du programme. À l'issue de la période de deux ans, une seconde ronde de données est collectée à des fins d'évaluation auprès de l'échantillon des 4 959 ménages[1].

La présidente et le ministre de la Santé vous chargent de superviser l'évaluation d'impact du PSAM et de formuler des recommandations quant à l'opportunité de l'étendre ou non à l'ensemble du pays. Dans le cas présent, vous devez répondre à la question suivante : *de combien le PSAM a-t-il réduit les dépenses de santé des ménages ruraux pauvres ?* Les enjeux sont importants. S'il s'avère que le PSAM a permis de réduire les dépenses de santé d'au moins neuf dollars, il sera élargi à tout le pays. Si, en revanche, l'objectif des neuf dollars n'a pas été atteint, vous recommanderez de ne pas étendre le programme.

Le premier « expert » en évaluation que vous consultez soutient que pour estimer l'impact du PSAM, il faut déterminer le changement dans les dépenses de santé des ménages inscrits au programme à travers le temps. Selon le consultant, puisque le PSAM couvre l'ensemble des dépenses de soins de santé primaires et des achats de médicaments, toute baisse des dépenses dans le temps peut être attribuée, pour l'essentiel, au PSAM. En vous fondant uniquement sur le sous-groupe des ménages inscrits, vous estimez les dépenses moyennes de santé lors de l'enquête de référence puis deux ans après la mise en œuvre du programme. Autrement dit, vous procédez à une évaluation avant-après. Le tableau 3.1 en présente les résultats.

Tableau 3.1 Cas 1—Impact du PSAM selon la méthode avant-après (comparaison de moyennes)

	Après	Avant	Différence	Stat. *t*
Dépenses de santé des ménages	7,8	14,4	–6,6	–28,9

Vous remarquez que les ménages inscrits au PSAM voient leurs dépenses directes de santé passer de 14,4 dollars avant l'introduction du PSAM à 7,8 dollars deux années plus tard, soit une baisse de 6,6 dollars (ou 45 %) sur la période. Comme le montre la valeur de la statistique *t*, la différence entre les dépenses de santé avant et après la mise en œuvre du programme est *statistiquement significative*, autrement dit la probabilité que l'impact estimé soit statistiquement nul est très faible.

Même si la comparaison avant-après porte sur le même groupe de ménages, vous craignez que certains facteurs aient pu évoluer au cours du temps et exercer un impact sur les dépenses de santé. Par exemple, plusieurs interventions dans le domaine de la santé ont eu lieu simultanément dans les villages concernés par le programme pilote. Par ailleurs, il est possible que les dépenses des ménages aient été affectées par la crise financière qu'a récemment connue le pays. Face à ces craintes, le consultant propose une *analyse de régression* plus sophistiquée censée permettre de tenir compte de tous ces facteurs externes. Les résultats de cette analyse sont présentés dans le tableau 3.2.

La régression linéaire analyse comment les dépenses de santé varient selon une variable binaire (0-1) pour laquelle le 0 correspond à l'observation au moment de l'enquête de référence et le 1 à l'observation au moment de l'enquête de suivi. La régression linéaire multivariée permet en plus de *contrôler pour* ou de *maintenir constantes* d'autres caractéristiques observées des ménages de l'échantillon, par exemple des indicateurs de fortune (actifs), la composition des ménages, etc. Vous notez que la régression linéaire simple est équivalente à la simple différence avant-après constatée pour les dépenses de santé (une réduction de 6,59 dollars). En contrôlant pour les autres facteurs dans vos données, vous obtenez un résultat semblable, à savoir une baisse de 6,65 dollars.

Tableau 3.2 Cas 1—Impact du PSAM selon la méthode avant-après (analyse de régression)

	Régression linéaire	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–6,59** (0,22)	–6,65** (0,22)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

QUESTION 1

A. Au vu des résultats pour le cas 1, le PSAM doit-il être élargi à l'échelle nationale ?

B. Cette analyse tient-elle compte de tous les facteurs qui peuvent influencer les dépenses de santé au fil du temps ?

Contrefactuel contrefait 2 : comparaison entre participants et non participants

La comparaison entre des unités bénéficiaires du programme et des unités n'en bénéficiant pas (« avec-sans ») constitue un autre contrefactuel contrefait. Prenons l'exemple d'un programme de formation professionnelle destiné à des jeunes sans emploi. Imaginons que deux ans après le lancement du programme, une évaluation tente d'estimer l'impact du programme sur les revenus en comparant les revenus moyens d'un groupe de jeunes ayant participé au programme aux revenus de ceux qui n'y ont pas participé. Supposons que les jeunes ayant participé au programme aient un revenu deux fois supérieur à celui des jeunes n'ayant pas participé.

Comment ces résultats doivent-ils être interprétés ? Dans ce cas, l'estimation du contrefactuel provient des revenus des personnes ayant décidé de ne pas participer au programme. Cependant, les deux groupes de jeunes ont de fortes chances de présenter des différences fondamentales. Ceux qui ont décidé d'intégrer le programme sont peut-être motivés par la perspective d'améliorer leurs conditions de vie et espèrent peut-être beaucoup bénéficier de la formation. À l'inverse, ceux qui ont préféré ne pas participer au programme sont peut-être des personnes découragées qui n'attendent rien de ce genre de programme. Il est probable que ces deux groupes de jeunes n'auraient pas le même parcours professionnel et que leurs revenus seraient différents même si le programme de formation professionnelle n'avait pas existé.

Le groupe ayant décidé de ne pas participer au programme ne permet donc pas d'obtenir un contrefactuel convaincant. Si une différence de revenus est observée entre les deux groupes, il sera impossible de l'attribuer à la formation professionnelle, à une différence de motivation ou à une quelconque autre différence entre les deux groupes. Le fait que les individus les moins motivés préfèrent ne pas participer au programme de formation introduit donc un biais dans l'estimation de l'impact du programme[2]. Ce biais est appelé « biais de sélection ». Dans cet exemple, si les jeunes gens ayant participé au programme avaient des revenus supérieurs même en l'absence du programme, le biais de sélection serait positif ; autrement dit, nous aurions surestimé l'impact du programme de formation professionnelle sur les revenus en comparant simplement les bénéficiaires aux non-bénéficiaires.

Concept clé :

Un biais de sélection apparait lorsque les raisons pour lesquelles une personne participe à un programme sont corrélées aux résultats. Ce biais se produit généralement lorsque le groupe de comparaison n'est pas éligible au programme ou décide de ne pas y participer.

Comparaison entre participants et non participants au Programme de subvention de l'assurance maladie (PSAM)

Suite à la réflexion suscitée par la comparaison avant-après au sein de votre équipe d'évaluation, vous êtes conscients que de nombreux facteurs variables dans le temps restent susceptibles d'expliquer la baisse des dépenses de santé (en particulier, le ministère des Finances craint que la récente crise financière ait joué un rôle dans

les dépenses de santé des ménages, facteur qui pourrait expliquer les changements observés). Un autre consultant suggère qu'il serait plus approprié d'estimer le contrefactuel à partir de l'enquête réalisée après l'intervention, c'est-à-dire deux ans après le lancement du programme. Le consultant fait remarquer, à juste titre, que sur les 4 959 ménages de l'échantillon de référence, seuls 2 907 ont effectivement participé au programme. Autrement dit, environ 41 % des ménages de l'échantillon n'ont pas été couverts par le PSAM. Il avance en outre que les ménages d'une même localité sont exposés à la même offre de soins et confrontés aux mêmes conditions économiques. Selon lui, les résultats mesurés après l'intervention auprès du groupe non inscrits au PSAM permettraient donc de tenir compte de nombreux facteurs contextuels qui touchent tous les ménages, qu'ils soient ou non inscrits au programme.

Vous décidez donc de calculer les dépenses de santé moyennes après l'intervention pour, d'une part, les ménages ayant participé au programme et, d'autre part, ceux qui n'y ont pas participé. Les observations recueillies sont présentées dans le tableau 3.3.

En vous fondant sur les dépenses de santé moyennes des ménages non inscrits pour élaborer le contrefactuel, vous aboutissez à la conclusion que le programme a permis de réduire les dépenses de santé moyennes d'environ 14 dollars. En discutant de ce résultat avec le consultant, vous soulevez la question de savoir si les ménages ayant choisi de ne pas participer au programme peuvent différer systématiquement de ceux qui ont choisi d'y participer. Par exemple, il est possible que les ménages ayant intégré le PSAM s'attendaient à une hausse de leurs dépenses de santé ou soient mieux informés sur le programme, ou encore qu'il s'agisse de personnes davantage préoccupées par la santé de leur famille. Il pourrait aussi s'agir de ménages plus pauvres en moyenne que ceux qui n'ont pas participé au PSAM, qui visait les ménages pauvres. Votre consultant affirme qu'une analyse de régression permet de prendre en compte les éventuelles différences entre les deux groupes. En tenant compte de toutes les caractéristiques de l'ensemble des ménages pour lesquels des données ont été recueillies, le consultant aboutit aux résultats présentés dans le tableau 3.4.

Tableau 3.3 Cas 2—Impact du PSAM selon la méthode avec-sans (comparaison de moyennes)

	Participants	Non Participants	Différence	Stat. de *t*
Dépenses de santé des ménages	7,8	21,8	–13,9	–39,5

Tableau 3.4 Cas 2—Impact du PSAM selon la méthode avec-sans (analyse de régression)

	Régression linéaire	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–13,9** (0,35)	–9,4** (0,32)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

Avec une régression linéaire simple des dépenses de santé sur un indicateur binaire (participation ou non d'un ménage au programme), vous obtenez un impact estimé de 13,90 dollars, autrement dit, vous estimez que le programme a diminué les dépenses de santé moyenne de 13,90 dollars. En revanche, si l'on tient compte de toutes les autres caractéristiques de la population de l'échantillon, la réduction des dépenses de santé des ménages ayant participé au programme s'élève à 9,40 dollars par an.

QUESTION 2

A. Au vu de ces résultats pour le cas 2, le PSAM doit-il être élargi à l'échelle nationale ?

B. Peut-on considérer que cette analyse tient compte de tous les facteurs susceptibles d'engendrer des différences entre les dépenses de santé des deux groupes ?

Notes

1. Nous supposons ici une attrition nulle de l'échantillon entre les deux enquêtes, autrement dit aucun ménage ne quitte l'échantillon. Cette hypothèse n'est pas réaliste pour la plupart des enquêtes sur les ménages. Dans les faits, les familles qui déménagent ne peuvent parfois pas être suivies et certains ménages se dissolvent. Le chapitre 12 discute des problèmes d'attrition en plus de détails.
2. Pour donner un autre exemple, si les jeunes qui pensent tirer davantage profit du programme de formation sont plus enclins à participer à la formation (par exemple parce qu'ils pensent que celle-ci leur permettra d'obtenir des salaires plus élevés), nous comparerions alors un groupe d'individus qui anticipent un revenu plus élevé avec un groupe d'individus qui n'anticipaient pas un revenu plus élevé.

CHAPITRE 4

Méthodes de sélection aléatoire

Le chapitre précédent a passé en revue deux approches (la comparaison avant-après et la comparaison avec-sans) communément utilisées pour l'élaboration de contrefactuels, mais présentant de forts risques de biais. Nous allons maintenant aborder une série de méthodes qui permettent d'évaluer l'impact d'un programme de manière plus fiable. Comme nous le verrons, l'exercice n'est pas aussi simple qu'il y paraît. La plupart des programmes sont conçus et mis en œuvre dans un environnement complexe et évoluant dans lequel de nombreux facteurs peuvent influencer les résultats tant pour les participants au programme que pour les non participants. Les sécheresses, les tremblements de terre, les transitions de gouvernement, les changements des politiques locales et internationales sont autant d'éléments inhérents au monde dans lequel nous vivons ; en tant qu'évaluateurs, nous voulons nous assurer que l'évaluation de l'impact d'un programme soit valide malgré ces nombreux facteurs.

Comme nous le verrons dans cette partie du manuel, les règles de sélection des bénéficiaires d'un programme constituent le paramètre clef pour choisir une méthode d'évaluation d'impact. Nous pensons que dans la plupart des cas, les méthodes d'évaluation doivent être adaptées aux règles opérationnelles d'un programme (avec quelques ajustements ici et là) et non le contraire. Toutefois, nous partons aussi de la prémisse que *tous les programmes sociaux doivent comprendre des règles de sélection des bénéficiaires justes et transparentes*. L'une des règles les plus justes et les plus transparentes pour allouer des ressources limitées parmi des populations dans le même besoin consiste à donner à toute personne éligible une chance égale de bénéficier du programme. À cet effet, une manière de faire consiste à procéder à un tirage au sort. Dans ce chapitre, nous allons examiner plusieurs *méthodes de sélection aléatoire* ; celles-ci consistent à effectuer des tirages au sort pour désigner lesquelles des

unités également éligibles participeront à un programme et lesquelles n'y participeront pas. Ces méthodes de sélection aléatoire permettent non seulement aux gestionnaires de programme de disposer de règles justes et transparentes pour distribuer des ressources limitées parmi *des populations dans le même besoin*, mais constituent aussi les méthodes les plus solides pour évaluer l'impact d'un programme.

Les méthodes de sélection aléatoire peuvent souvent découler des règles opérationnelles d'un programme. Dans de nombreux programmes, la population des participants visés, c'est-à-dire le groupe de toutes les unités que le programme voudrait atteindre, est plus grande que le nombre de participants que le programme peut servir à un moment donné. Par exemple, en une année, un programme d'éducation peut fournir du matériel scolaire et un curriculum amélioré à 500 écoles sur les milliers d'écoles éligibles que compte un pays. Autre exemple, un programme d'emploi pour les jeunes peut avoir pour objectif de toucher 2 000 jeunes chômeurs durant sa première année d'opération, même s'il y a des dizaines de milliers de jeunes chômeurs dans le pays que le programme viserait ultimement à incorporer. Il y a de multiples raisons qui font que les programmes peuvent être dans l'incapacité de servir l'ensemble de leur population cible. Des contraintes budgétaires peuvent empêcher le programme de couvrir l'ensemble des unités éligibles dès son lancement. Même si les budgets sont suffisants pour servir un nombre illimité de participants, les capacités peuvent manquer pour que le programme incorpore l'ensemble de la population cible au même moment. Dans l'exemple du programme de formation professionnelle destiné aux jeunes, le nombre de jeunes chômeurs désirant intégrer une formation peut être supérieur au nombre de places disponibles dans les écoles techniques durant la première année de mise en œuvre du programme, ce qui limite le nombre de jeunes qui peuvent participer au programme.

Dans la réalité, la plupart des programmes sont tributaires de contraintes budgétaires ou opérationnelles qui les empêchent d'atteindre toute la population cible au même moment. Dans le cas où le nombre de personnes éligibles au programme est supérieur au nombre de places offertes, les gestionnaires doivent définir un mécanisme d'allocation des bénéfices du programme. Autrement dit, quelqu'un doit décider qui pourra participer au programme et qui ne pourra pas y participer. Les bénéfices peuvent être alloués selon la règle du « premier arrivé, premier servi » ou sur la base de certaines caractéristiques observées (par exemple les femmes et les enfants d'abord, ou encore les localités les plus pauvres d'abord) ; la sélection peut aussi s'effectuer selon des caractéristiques non observées (par exemple laisser les personnes intégrer le programme en fonction de leur motivation ou de leurs connaissances) ou même par tirage au sort.

Assignation aléatoire du traitement

Lorsqu'un programme est distribué de manière aléatoire parmi une population éligible nombreuse, il est possible de générer un contrefactuel solide considéré comme l'étalon-or en matière d'évaluation d'impact. L'assignation aléatoire du traitement

repose, pour l'essentiel, sur l'utilisation d'un tirage au sort pour désigner les bénéficiaires du programme[1] parmi une population d'unités tout aussi éligibles les unes que les autres. La probabilité d'être sélectionnée est alors la même pour toutes les unités éligibles (une personne, un ménage, une communauté, une école, un hôpital, etc.)[2].

Avant d'évoquer l'application pratique de l'assignation aléatoire et les raisons pour lesquelles cette méthode permet d'obtenir un contrefactuel solide, examinons pourquoi l'assignation aléatoire est considérée comme un moyen juste et transparent d'allouer des ressources limitées. Une fois qu'une population cible a été définie (par exemple, les ménages vivant au-dessous du seuil de pauvreté, les enfants de moins de cinq ans ou encore les écoles situées en milieu rural), l'assignation aléatoire peut être considérée comme une règle juste, car elle assure au gestionnaire de programme que toute personne ou unité éligible possède la même chance de participer au programme et qu'aucun critère arbitraire ou subjectif, ni aucun favoritisme ou autre pratique inéquitable n'interviennent. Quand la demande est supérieure à l'offre, l'assignation aléatoire est une règle facilement justifiable par les gestionnaires de programme et facilement comprise par les principales parties prenantes. Lorsque la sélection des bénéficiaires s'effectue selon un processus transparent et vérifiable, la règle de l'assignation aléatoire ne peut pas être aisément manipulée ; elle protège donc les gestionnaires de programme d'éventuelles accusations de favoritisme ou de corruption. L'assignation aléatoire présente en ce sens des avantages au-delà de sa seule utilité pour l'évaluation d'impact. De nombreux programmes ont d'ailleurs recours à des tirages au sort afin de sélectionner des participants à partir d'un groupe d'individus éligibles, et ce en raison des avantages de cette technique pour la gestion et la gouvernance des programmes[3].

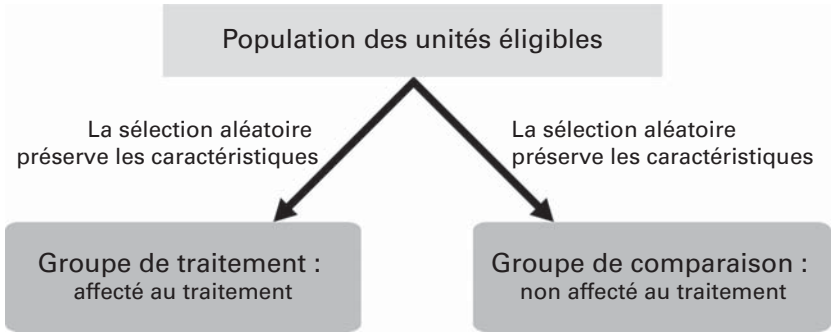
Pourquoi l'assignation aléatoire produit-elle une excellente estimation du contrefactuel ?

Comme nous l'avons souligné ci-dessus, un groupe de comparaison idéal est en tout point similaire au groupe de traitement à la seule différence qu'il ne participe pas au programme à évaluer. La sélection aléatoire des unités qui bénéficieront du traitement et de celles qui feront partie des groupes de comparaison génère deux groupes ayant une forte probabilité d'être statistiquement identiques, pour autant que le nombre d'unités auxquelles est appliqué le processus d'assignation aléatoire soit assez important. Plus précisément, avec un nombre suffisamment important d'observations, le processus d'assignation aléatoire permet de constituer des groupes *dont toutes les caractéristiques moyennes sont statistiquement équivalentes*. À leur tour, ces moyennes tendent vers la moyenne de la population dont elles sont issues[4].

La figure 4.1 illustre pourquoi l'assignation aléatoire fournit un groupe de comparaison statistiquement équivalent au groupe de traitement. Supposons que la population des unités éligibles (participants potentiels) comprenne 1 000 personnes dont la moitié a été sélectionnée de manière aléatoire pour faire partie du groupe de traitement, l'autre moitié constituant le groupe de comparaison. Par exemple, imaginons écrire les noms des 1 000 personnes sur de petits bouts de papier, les mettre dans une urne et tirer au sort 500 noms. S'il a été décidé que les 500 premiers noms tirés au sort feront partie du groupe de traitement, nous obtiendrons alors un groupe de traitement (les 500 premiers noms tirés) et un groupe de comparaison (les 500 noms restant dans l'urne), tous deux constitués de manière aléatoire.

Imaginons maintenant que sur les 1 000 personnes, 40 % soient des femmes. Comme les noms ont été sélectionnés au hasard, environ 40 % des 500 noms tirés de l'urne seront aussi des femmes. Si 20 % des 1 000 personnes ont des yeux bleus, la proportion d'yeux bleus sera à peu près la même dans le groupe de traitement et dans le groupe de comparaison. En général, si la population des unités éligibles est suffisamment nombreuse, les caractéristiques de la population se transmettront au groupe de traitement et au groupe de comparaison. Si des caractéristiques observables comme le genre ou la couleur des yeux se transmettent aux deux groupes, il semble logique de considérer que des caractéristiques plus difficiles à observer (des variables non observées) comme la motivation, les préférences ou les traits de personnalité, se transmettront aussi de manière équivalente de la population au groupe de comparaison et au groupe de traitement. Le groupe de traitement et le groupe de comparaison constitués par assignation aléatoire seront donc similaires à la population de référence non seulement sur le plan des caractéristiques observées, mais aussi des caractéristiques non observées. Par exemple, il est difficile d'observer ou de mesurer l' « amabilité », mais si l'on sait que les personnes aimables représentent 20 % de la population des unités éligibles, le groupe de traitement et le groupe

Figure 4.1 Caractéristiques des groupes constitués par assignation aléatoire du traitement

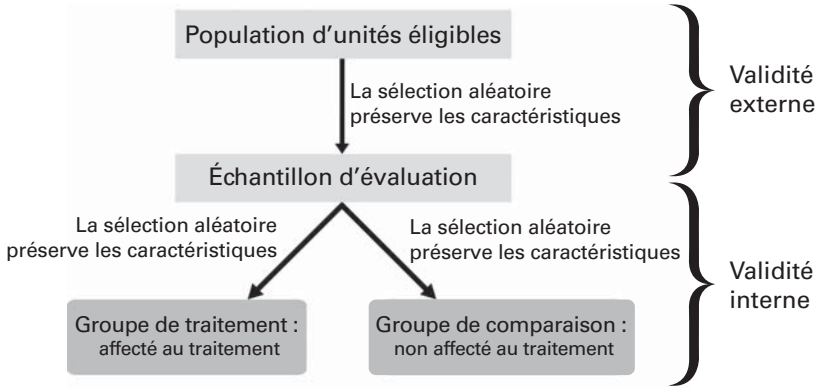


de comparaison comprendront la même proportion de personnes dotées de cette caractéristique. L'assignation aléatoire permet de garantir que le groupe de traitement et le groupe de comparaison seront en moyenne en tout point similaire tant au niveau des caractéristiques observées et non observées.

Dans le cadre d'une évaluation d'impact, l'utilisation de l'assignation aléatoire pour définir le groupe de traitement et le groupe de comparaison garantit en théorie que les groupes sont équivalents. La collecte de données de base pour un échantillon d'évaluation permet de vérifier empiriquement cette hypothèse, en s'assurant qu'il n'existe pas de différence systématique entre les caractéristiques observées des groupes de traitement et de comparaison avant que le programme ne débute. Dans ce cas, comme les deux groupes sont identiques au départ et sont exposés aux mêmes facteurs contextuels externes au cours du temps, toute différence observée entre les résultats des deux groupes après le lancement du programme peut être attribuée au programme. En d'autres termes, le groupe de comparaison permet de *contrôler* pour tous les autres facteurs qui peuvent potentiellement expliquer le résultat. Nous pouvons alors être sûrs que l'estimation de l'impact moyen obtenu par la différence entre le résultat observé dans le groupe de traitement (la moyenne des résultats pour le groupe de traitement constitué par assignation aléatoire) et l'estimation du contrefactuel (la moyenne des résultats pour le groupe de comparaison également constitué par assignation aléatoire) représente le véritable impact du programme. En effet, le processus de formation des groupes permet d'écarter tous les autres facteurs, observés ou non, qui auraient pu constituer une explication plausible de la différence des résultats entre les deux groupes.

La figure 4.1 suppose que toutes les unités de la population éligible sont réparties soit dans le groupe de traitement, soit dans le groupe de comparaison. Dans certains cas, il n'est pas nécessaire d'inclure toutes les unités de la population éligible dans le travail d'évaluation. Par exemple, si la population des unités éligibles est constituée d'un million de mères et que nous cherchons à évaluer l'efficacité de transferts monétaires sur la probabilité que ces mères fassent vacciner leurs enfants, il peut être suffisant de répartir un échantillon représentatif, par exemple de 1 000 personnes, entre le groupe de traitement et le groupe de comparaison. La figure 4.2 illustre ce processus. Par la même logique que ci-dessus, la sélection d'un échantillon aléatoire à partir de la population des unités éligibles permet de préserver les caractéristiques de la population dans l'échantillon. La sélection aléatoire des groupes de traitement et de comparaison à partir de l'échantillon préservera à son tour les caractéristiques de la population dans chaque groupe.

Figure 4.2 Échantillonnage aléatoire et assignation aléatoire du traitement



Validité interne et validité externe

Concept clé :
Une évaluation possède une validité interne si elle est fondée sur un groupe de comparaison valide.

Les étapes décrites ci-dessus pour l'assignation aléatoire du traitement permettent de garantir la validité tant interne qu'externe de l'évaluation d'impact (figure 4.2) pour autant que l'échantillon d'évaluation soit de taille suffisante.

La validité interne signifie que l'impact estimé ne peut pas être influencé par des facteurs autres que le programme, autrement dit, que le groupe de comparaison constitue un contrefactuel valable permettant d'estimer l'impact réel du programme. Pour rappel, l'assignation aléatoire permet de former un groupe de comparaison statistiquement équivalent au groupe de traitement avant que le programme ne débute. Après le lancement du programme, le groupe de comparaison est soumis aux mêmes facteurs externes que le groupe de traitement à la seule différence qu'il n'est pas exposé au programme. Dès lors, si des différences apparaissent entre le groupe de comparaison et le groupe de traitement, elles ne peuvent être attribuées qu'au programme. Autrement dit, la validité interne d'une évaluation d'impact est assurée par le processus d'*assignation aléatoire du traitement*.

Concept clé :
Une évaluation possède une validité externe si l'échantillon d'évaluation est représentatif de la population des unités éligibles. Les résultats obtenus pour l'échantillon peuvent alors être généralisés à l'ensemble de la population des unités éligibles.

La validité externe signifie que l'impact estimé pour l'échantillon d'évaluation peut être généralisé à toute la population des unités éligibles. Pour que cela soit possible, il faut que l'échantillon d'évaluation soit représentatif de la population des unités éligibles ; dans les faits, cela suppose que l'échantillon soit constitué à partir de la population en utilisant une méthode d'*échantillonnage aléatoire*[5].

Nous avons évoqué deux types de sélection aléatoire : la première à des fins d'échantillonnage (pour la validité externe) et la seconde en tant que méthode d'évaluation d'impact (pour la validité interne). Une évaluation d'impact peut produire des estimations ayant une solide validité interne en utilisant une assignation aléatoire du traitement, mais si l'évaluation est effectuée sur un échantillon sélectionné

de manière non aléatoire, l'impact estimé peut ne pas être généralisé à l'ensemble de la population des unités éligibles. De même, si l'évaluation est fondée sur un échantillon sélectionné de manière aléatoire, mais que le traitement n'est pas distribué de manière aléatoire, l'échantillon sera certes représentatif, mais le groupe de comparaison peut ne pas être valide.

Quand utiliser l'assignation aléatoire ?

Dans la pratique, l'assignation aléatoire peut être considérée pour tout programme pour lequel la demande excède l'offre, c'est-à-dire lorsque le nombre de participants potentiels dépasse les capacités du programme à un moment donné et que ce programme doit être graduellement élargi. Dans d'autres cas, une assignation aléatoire se justifie à des fins d'évaluation même si les ressources du programme sont illimitées. Par exemple, les autorités peuvent recourir à l'assignation aléatoire pour éprouver de nouveaux programmes potentiellement coûteux dont les effets recherchés et indésirables restent méconnus. Dans de telles circonstances, l'assignation aléatoire peut être utilisée durant la phase d'évaluation pilote pour déterminer avec précision les effets du programme avant de l'élargir à une population plus importante.

L'assignation aléatoire constitue une méthode d'évaluation d'impact adéquate dans deux cas fréquents :

1. *Si le nombre d'unités éligibles est supérieur au nombre de places disponibles dans le programme*. Si la demande dépasse l'offre, un tirage au sort peut être effectué pour définir le groupe qui bénéficiera du programme parmi la population éligible. Dans ce cas, toutes les unités de la population ont la même chance d'être sélectionnées. Le groupe des unités tirées au sort constitue le groupe de traitement et le reste de la population, qui ne bénéficiera pas du programme, le groupe de comparaison. Aussi longtemps que des contraintes de ressources empêchent d'étendre le programme à l'ensemble de la population, les groupes de comparaison peuvent être maintenus pour mesurer l'impact du programme à court, moyen et long terme. Dans ces conditions, il n'y a pas de dilemme éthique à garder indéfiniment un groupe de comparaison puisqu'une partie de la population ne peut de toute façon pas être couverte par le programme.

   Par exemple, supposons que le ministère de l'Éducation d'un pays souhaite doter les écoles publiques de bibliothèques, mais que le budget mis à disposition par le ministère des Finances ne permet de couvrir qu'un tiers des écoles. Si le ministère de l'Éducation souhaite donner une chance égale d'obtenir une bibliothèque à chacune des écoles publiques, il peut procéder à un tirage au sort au cours duquel chaque école a une chance égale (c'est-à-dire une chance sur trois) d'être sélectionnée. Les écoles tirées au sort seront dotées d'une nouvelle bibliothèque

et constitueront le groupe de traitement, et les écoles restantes, c'est-à-dire les deux tiers des écoles totales, qui n'auront pas de bibliothèque formeront le groupe de comparaison. À moins que des fonds supplémentaires ne soient alloués au programme de bibliothèques, il restera un groupe d'écoles qui ne pourront pas être dotées de bibliothèque dans le cadre du programme et qui pourront servir de groupe de comparaison pour estimer le contrefactuel.

2. *Lorsqu'un programme doit être progressivement étendu pour couvrir l'ensemble de la population éligible.* Quand un programme est graduellement mis en œuvre, la sélection aléatoire de l'ordre dans lequel les participants bénéficieront du programme donne à chaque unité éligible une chance égale de recevoir le traitement à la première phase ou à une phase ultérieure du programme. Tant que le « dernier » groupe n'aura pas intégré le programme, il constituera le groupe de comparaison servant à estimer le contrefactuel pour les groupes ayant déjà été soumis au traitement.

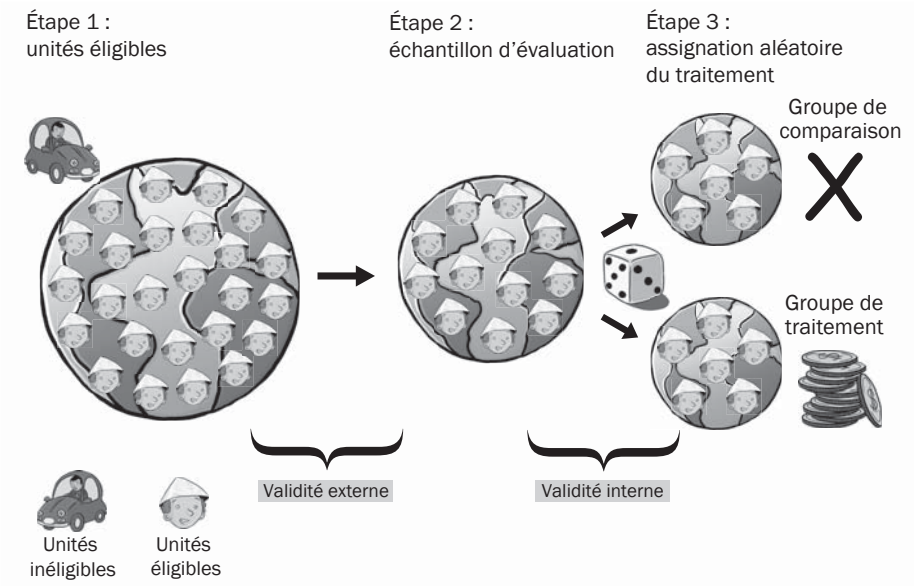
   Imaginons que le ministère de la Santé souhaite former les 15 000 infirmières du pays à un nouveau protocole, mais qu'il faille trois années pour toutes les former. À des fins d'évaluation d'impact, le ministère peut sélectionner de manière aléatoire un premier tiers d'infirmières qui suivront la formation la première année, un second tiers la deuxième année et un dernier tiers la troisième année. Pour évaluer l'impact du programme de formation une année après son lancement, le groupe d'infirmières ayant bénéficié de la formation la première année constituera le groupe de traitement, et le groupe qui a été choisi aléatoirement pour suivre la formation la troisième année constituera le groupe de comparaison puisqu'il n'aura pas encore bénéficié de la formation.

Comment réaliser l'assignation aléatoire ?

Nous avons évoqué la méthode de l'assignation aléatoire et les raisons pour lesquelles elle permet de créer un groupe de comparaison valable. Nous allons maintenant examiner les étapes à respecter pour réaliser l'assignation aléatoire d'un traitement. La figure 4.3 illustre ce processus.

La première étape consiste à définir les unités éligibles au programme. Selon les programmes, l'unité peut être une personne, un centre de santé, une école ou même tout un village ou une municipalité. La population des unités éligibles comprend toutes les unités pour lesquelles vous cherchez à déterminer l'impact du programme. Par exemple, dans le cadre d'une évaluation d'un programme de formation des instituteurs d'écoles primaires en zones rurales, les professeurs du secondaire et ceux des écoles primaires en milieu urbain ne feront pas partie de la population des unités éligibles.

Figure 4.3 Étapes de l'assignation aléatoire du traitement

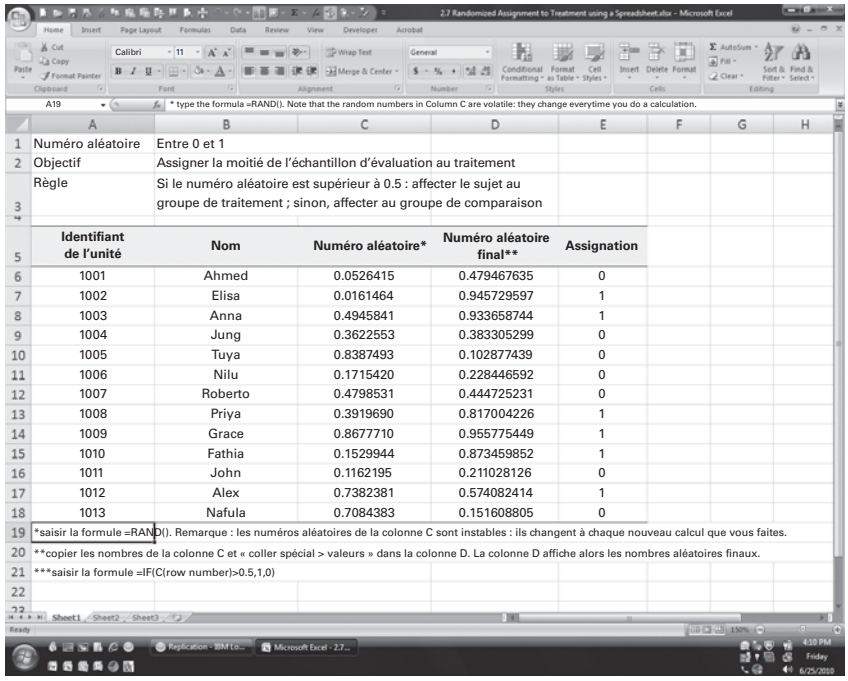


Une fois que la population des unités éligibles est définie, il faudra comparer la taille du groupe avec le nombre d'observations requises pour l'évaluation. Ce nombre est déterminé par des calculs de puissance et dépend du type de questions auxquelles vous voulez répondre (voir chapitre 11). Si la population éligible est peu nombreuse, il peut être nécessaire d'inclure toutes les unités éligibles dans l'évaluation. À l'inverse, s'il y a plus d'unités éligibles que nécessaire pour l'évaluation, la deuxième étape consistera à sélectionner un échantillon d'unités à partir de la population pour élaborer un échantillon d'évaluation. Cette deuxième étape vise essentiellement à limiter les coûts de collecte des données. Si les données fournies par les systèmes de suivi existants peuvent être utilisées pour effectuer l'évaluation et que ces systèmes couvrent la population des unités éligibles, la création d'un échantillon d'évaluation distinct n'est pas nécessaire. Par contre, imaginons que vous devez collecter des données détaillées sur les connaissances pédagogiques de plusieurs dizaines de milliers de professeurs dans toutes les écoles publiques du pays. Réaliser des entretiens avec chaque professeur risque fort d'être impossible ; mais un échantillon de 1 000 professeurs travaillant dans 100 écoles différentes peut être suffisant. Si l'échantillon est représentatif de l'ensemble de la population des enseignants des écoles publiques, les résultats de l'évaluation resteront généralisables à l'ensemble des professeurs et écoles publiques du pays. Recueillir des données auprès d'un échantillon de 1 000 professeurs sera bien évidemment moins coûteux que de s'entretenir avec tous les professeurs des écoles publiques du pays.

Enfin, la troisième étape consistera à former le groupe de traitement et le groupe de comparaison à partir des unités de l'échantillon d'évaluation. En ce sens, vous devez tout d'abord établir une règle de répartition des participants sur la base de nombres aléatoires. Par exemple, pour affecter 40 des 100 unités de l'échantillon d'évaluation au groupe de traitement, vous pourrez établir la règle selon laquelle les 40 unités qui ont reçu les numéros aléatoires les plus élevés constitueront le groupe de traitement et que les autres unités formeront le groupe de comparaison. Vous attribuerez donc un numéro aléatoire à chaque unité de l'échantillon d'évaluation à l'aide d'une feuille de calcul ou d'un logiciel statistique spécialisé (figure 4.4). À partir de la règle préalablement établie, vous pourrez ensuite constituer un groupe de comparaison et un groupe de traitement. Il est important de convenir de la règle avant d'utiliser le logiciel qui attribuera les nombres aléatoires aux unités. Dans le cas contraire, vous pourriez être tenté de choisir la règle en fonction des nombres aléatoires observés, ce qui invaliderait automatiquement le processus d'assignation aléatoire.

La logique sous-jacente au processus automatisé est la même que celle qui prévaut lors d'un tirage à pile ou face ou du tirage au hasard d'un nom d'un chapeau : dans tous les cas, il s'agit de laisser le hasard déterminer à quel groupe (groupe de traitement ou

Figure 4.4 Assignation aléatoire du traitement avec utilisation d'une feuille de calcul



	A	B	C	D	E	F	G	H
1	Numéro aléatoire	Entre 0 et 1						
2	Objectif	Assigner la moitié de l'échantillon d'évaluation au traitement						
3	Règle	Si le numéro aléatoire est supérieur à 0.5 : affecter le sujet au groupe de traitement ; sinon, affecter au groupe de comparaison						
5	**Identifiant de l'unité**	**Nom**	**Numéro aléatoire***	**Numéro aléatoire final****	**Assignation**			
6	1001	Ahmed	0.0526415	0.479467635	0			
7	1002	Elisa	0.0161464	0.945729597	1			
8	1003	Anna	0.4945841	0.933658744	1			
9	1004	Jung	0.3622553	0.383305299	0			
10	1005	Tuya	0.8387493	0.102877439	0			
11	1006	Nilu	0.1715420	0.228446592	0			
12	1007	Roberto	0.4798531	0.444725231	0			
13	1008	Priya	0.3919690	0.817004226	1			
14	1009	Grace	0.8677710	0.955775449	1			
15	1010	Fathia	0.1529944	0.873459852	1			
16	1011	John	0.1162195	0.211028126	0			
17	1012	Alex	0.7382381	0.574082414	1			
18	1013	Nafula	0.7084383	0.151608805	0			
19	*saisir la formule =RAND(). Remarque : les numéros aléatoires de la colonne C sont instables : ils changent à chaque nouveau calcul que vous faites.							
20	**copier les nombres de la colonne C et « coller spécial > valeurs » dans la colonne D. La colonne D affiche alors les nombres aléatoires finaux.							
21	***saisir la formule =IF(C(row number)>0.5,1,0)							

groupe de comparaison) appartient chaque unité. Quand l'assignation aléatoire doit se faire en public, des méthodes plus « artisanales » peuvent être utilisées. Les exemples suivants supposent que l'unité de sélection aléatoire est une personne :

1. Si vous souhaitez placer 50 % des personnes dans le groupe de traitement et 50 % des personnes dans le groupe de comparaison, vous pouvez procéder à un tirage à pile ou face pour chacune d'elles. Vous devez alors décider préalablement si les personnes tirées à pile ou celles tirées à face formeront le groupe de traitement.

2. Si vous voulez qu'un tiers de l'échantillon d'évaluation constitue le groupe de traitement, vous pouvez lancer un dé pour chaque personne. Vous devez toutefois décider d'une règle d'attribution au préalable. Par exemple, si le dé tombe sur le un ou sur le deux, la personne fera partie du groupe de traitement, alors que si c'est le trois, le quatre, le cinq ou le six qui sort, la personne fera partie du groupe de comparaison. Vous lancerez le dé une fois pour chaque personne faisant partie de l'échantillon d'évaluation, et la personne sera ensuite affectée au groupe de traitement ou de comparaison en fonction du numéro qui sort.

3. Inscrivez les noms de toutes les personnes sur des papiers de taille et de forme identiques. Pliez les papiers de manière à ce que les noms soient invisibles et mélangez-les dans un chapeau ou tout autre récipient. Avant le tirage au sort, fixez une règle en définissant le nombre de papiers qui seront tirés au sort et si les noms tirés seront affectés au groupe de traitement ou au groupe de comparaison. Dès que la règle est établie, demandez à l'une des personnes présentes (quelqu'un d'impartial, par exemple un enfant) de tirer autant des papiers jusqu'à ce que le nombre de participants dans le groupe de traitement soit atteint.

Qu'il s'agisse d'un tirage au sort en public, d'un lancer de dé ou de nombres aléatoires générés par un programme informatique, il est important de documenter le processus pour en assurer la transparence. À cet effet, il convient tout d'abord que la règle ait été préalablement convenue et communiquée aux témoins et participants. Il faut ensuite se tenir à cette règle lors du tirage au sort et être en mesure de démontrer que le processus est effectivement réalisé au hasard. Dans le cas de tirages au sort ou de lancers de dé, il est possible de filmer le processus ; si un programme informatique a été utilisé pour tirer des nombres aléatoires, il convient de sauvegarder un registre de vos calculs afin que des auditeurs puissent, le cas échéant, les répliquer[6].

À quel niveau réaliser l'assignation aléatoire ?

L'assignation aléatoire peut s'effectuer au niveau de l'individu, du ménage, de la communauté ou de la région. En général, le niveau auquel s'effectue l'assignation aléatoire des unités au groupe de traitement ou au groupe de comparaison dépend de la manière selon laquelle le programme est mis en œuvre. Par exemple, si un programme de santé est mis en œuvre au niveau des cliniques, vous pourrez d'abord établir un échantillon aléatoire de cliniques et procéder, dans un second temps, à leur assignation aléatoire soit au groupe de traitement, soit au groupe de comparaison.

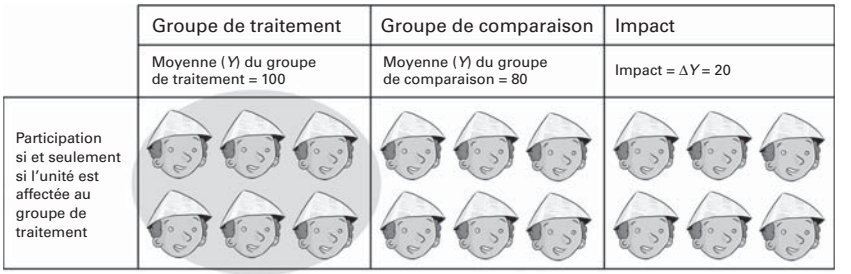
Quand l'assignation aléatoire est réalisée à un niveau plus élevé, par exemple au niveau des régions ou des provinces d'un pays, il peut être très difficile de procéder à une évaluation d'impact, car le nombre de régions et de provinces n'est généralement pas suffisant pour permettre de constituer des groupes de traitement et de comparaison adéquats. Par exemple, si un pays ne compte que six provinces, le groupe de traitement et le groupe de comparaison ne pourront pas compter plus de trois provinces chacun, ce qui est insuffisant pour garantir que les caractéristiques de ces deux groupes soient équilibrées.

À l'inverse, plus l'échelle diminue, par exemple en atteignant les personnes ou les ménages, et plus les risques d'effets de diffusion et de contamination augmentent[7]. Prenons l'exemple d'un programme consistant à fournir des médicaments vermifuges à des ménages. Si un ménage du groupe de traitement vit à proximité d'un ménage qui, lui, fait partie du groupe de comparaison, ce dernier peut bénéficier d'un effet de diffusion positif lié au traitement prodigué au ménage voisin. Le risque que le ménage du groupe de comparaison soit contaminé par son voisin se réduit. Il convient, dans un tel cas, de veiller à ce que les ménages du groupe de traitement soient physiquement suffisamment éloignés de ceux du groupe de comparaison pour éviter que les effets de diffusion n'affectent les résultats. Toutefois, plus la distance entre les ménages augmente, plus la mise en œuvre du programme et la réalisation des enquêtes seront coûteuses. En règle générale, si les risques d'effets de diffusion peuvent être raisonnablement écartés, l'idéal est de procéder à l'assignation aléatoire du traitement au niveau le plus bas auquel le programme est mis en œuvre afin de constituer des groupes de comparaison et de traitement comprenant le plus grand nombre possible d'unités. La question des effets de diffusion (ou effets de débordements) est abordée au chapitre 8.

Estimation d'impact avec assignation aléatoire

Une fois qu'un échantillon d'évaluation est formé et que le traitement est attribué de manière aléatoire, il est relativement facile d'estimer l'impact du programme. Après une certaine période de mise en œuvre du programme, il faudra mesurer les résultats pour les groupes de traitement et de comparaison. L'impact du programme correspond tout simplement à la différence entre le résultat moyen (Y) constaté pour le groupe de traitement et le résultat moyen (Y) observé pour le groupe de comparaison. Par exemple, à la figure 4.5, le résultat moyen est de 100 pour le groupe de traitement et de 80 pour le groupe de comparaison. L'impact du programme est donc de 20.

Figure 4.5 Estimation d'impact avec assignation aléatoire



Estimation d'impact du Programme de subvention de l'assurance maladie (PSAM) par assignation aléatoire

Revenons maintenant à notre exemple du PSAM (Programme de subvention de l'assurance maladie) et voyons ce que « l'assignation aléatoire » signifie dans ce cas-là. Rappelez-vous qu'il s'agit d'évaluer l'impact d'un programme à partir d'une phase pilote qui concerne 100 villages.

Après avoir mené deux évaluations d'impact avec des estimations du contrefactuel potentiellement biaisées (qui ont abouti à des recommandations opposées, voir chapitre 3), vous décidez de repartir à zéro et reconsidérez comment obtenir un contrefactuel plus précis. Après discussion avec votre équipe, vous êtes désormais convaincu que pour obtenir un contrefactuel valide, il faut identifier un groupe de villages de comparaison qui soient identiques en tout point aux 100 villages de traitement, à la seule différence que le premier groupe ne bénéficie pas du PSAM. Il s'avère que le PSAM a été lancé sous la forme d'un projet pilote et que les 100 villages participant à la première phase (villages de traitement) ont été désignés de manière aléatoire parmi l'ensemble des villages ruraux du pays. Vous notez que les 100 villages doivent donc, en moyenne, présenter les mêmes caractéristiques que la population générale des villages ruraux. Dans ce contexte, le contrefactuel peut être estimé de manière valide en mesurant les dépenses de santé des ménages éligibles dans les villages ne participant pas au PSAM.

Par chance, au moment de la réalisation des enquêtes de référence et de suivi, l'entreprise de sondage a recueilli des informations sur 100 villages ruraux supplémentaires qui n'ont pas été couverts par le PSAM lors de la phase pilote. Tout comme les villages de traitement, ces 100 villages ont été sélectionnés de manière aléatoire parmi la population des villages éligibles, ce qui signifie qu'ils présentent, en moyenne, les mêmes caractéristiques que toute la population des villages ruraux. La manière dont les deux groupes de villages ont été sélectionnés garantit donc qu'ils présentent des caractéristiques identiques, la seule différence étant que les 100 villages soumis au traitement bénéficient du PSAM, contrairement aux 100 autres villages, qui constituent le groupe de comparaison. Le traitement a été attribué de manière aléatoire.

Étant donné l'assignation aléatoire du traitement, vous êtes plutôt sûr qu'aucun facteur externe autre que le PSAM ne pourra expliquer les différences de résultats entre les villages de traitement et les villages de comparaison. Pour valider cette hypothèse, vous vérifiez que les ménages éligibles du groupe de traitement et de comparaison présentent bien les mêmes caractéristiques avant la mise en œuvre du programme (tableau 4.1).

Vous remarquez que les caractéristiques moyennes des ménages des deux groupes sont effectivement très proches. La seule différence statistiquement significative est le nombre d'années d'éducation du conjoint, mais cette différence est minime. Même si l'assignation aléatoire porte sur un grand échantillon, quelques rares différences entre les groupes de traitement et de comparaison peuvent subsister[8]. La validité du groupe de comparaison étant établie, vous estimez le contrefactuel par les dépenses de santé moyennes des ménages éligibles dans les 100 villages du groupe de comparaison (tableau 4.2).

Table 4.1 Cas 3— Comparabilité entre villages de traitement et villages de comparaison

Caractéristiques des ménages	Villages de traitement (N = 2 964)	Villages de comparaison (N = 2 664)	Différence	Stat. de *t*
Dépenses de santé (en dollars, par année et par personne)	14,48	14,57	–0,09	–0,39
Âge du chef du ménage (en années)	41,6	42,3	–0,7	–1,2
Âge du conjoint (en années)	36,8	36,8	0,0	0,38
Niveau d'éducation du chef du ménage (en années)	2,9	2,8	0,1	2,16*
Niveau d'éducation du conjoint (en années)	2,7	2,6	0,1	0,006
Le chef du ménage est une femme = 1	0,07	0,07	–0,0	–0,66
Autochtone = 1	0,42	0,42	0,0	0,21
Nombre de personnes dans le ménage	5,7	5,7	0,0	1,21
Présence d'une salle de bains = 1	0,57	0,56	0,01	1,04
Hectares de terre	1,67	1,71	–0,04	–1,35
Distance de l'hôpital (en km)	109	106	3	1,02

* Seuil de signification de 5 %.

Tableau 4.2 Cas 3— Impact du PSAM selon la méthode d'assignation aléatoire (comparaison des moyennes)

	Groupe de traitement	Groupe de comparaison	Différence	Stat. de *t*
Dépenses de santé des ménages observées lors de l'enquête de base	14,48	14,57	–0,09	–0,39
Dépenses de santé observées lors de l'enquête de suivi	7,8	17,9	–10,1**	–25,6

** Seuil de signification de 1 %.

Tableau 4.3 Cas 3— Impact du PSAM selon la méthode d'assignation aléatoire (analyse de régression)

	Régression linéaire	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–10,1** (0,39)	–10,0** (0,34)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

Vous disposez maintenant d'un contrefactuel valide et pouvez évaluer l'impact du PSAM en calculant la différence entre les dépenses de santé directes des ménages éligibles vivant dans les villages de traitement et l'estimation du contrefactuel. L'impact indique une baisse des dépenses de santé de 10,10 dollars sur deux ans. L'analyse de régression donne le même résultat, comme le montre le tableau 4.3.

Grâce à l'assignation aléatoire, nous pouvons être sûrs qu'aucun autre facteur systématiquement différent entre le groupe de traitement et le groupe de comparaison ne peut expliquer la différence des dépenses de santé. Les deux groupes de villages ont été exposés aux mêmes politiques et programmes nationaux au cours des deux années de la phase pilote du PSAM. Dans ces conditions, la raison la plus plausible pour expliquer que les ménages pauvres du groupe de traitement ont des dépenses inférieures à celles des ménages du groupe de comparaison est que les premiers ont bénéficié du programme d'assurance maladie, au contraire des seconds.

QUESTION 3

A. Pourquoi l'estimation d'impact à laquelle on aboutit avec la régression linéaire est-elle pratiquement inchangée en tenant compte d'autres facteurs ?

B. Au vu de ces résultats pour le cas 3, le PSAM doit-il être élargi à l'échelle nationale ?

L'assignation aléatoire en pratique

L'assignation aléatoire est souvent utilisée dans les études d'évaluation d'impact rigoureuses, tant pour les évaluations à grande échelle que de plus petite envergure. L'évaluation du programme Mexico Progresa (Schultz, 2004) est l'une des évaluations à grande échelle les plus connues utilisant l'assignation aléatoire (encadré 4.1).

Deux variations de l'assignation aléatoire

Nous allons maintenant aborder deux variations reposant sur les propriétés de l'assignation aléatoire : l'offre aléatoire et la promotion aléatoire du traitement.

Encadré 4.1 : Transferts monétaires conditionnels et éducation au Mexique

Le programme Progresa, qui s'appelle maintenant « Oportunidades », a été lancé en 1998 ; il propose un transfert monétaire aux mères pauvres vivant dans les régions rurales du Mexique à condition que leurs enfants soient présents à l'école et leur présence confirmée par l'enseignant. Ce programme social à grande échelle est l'un des premiers à avoir incorporé une évaluation d'impact rigoureuse. La méthode de l'assignation aléatoire a été utilisée pour permettre de déterminer les effets des transferts monétaires conditionnels sur un certain nombre de résultats, dont le taux de fréquentation scolaire.

Les bourses offertes aux enfants de la troisième[9] à la neuvième[10] année représentent entre 50 % et 75 % des frais de scolarité et sont attribuées pour une période de trois ans. Les communautés et les ménages éligibles pour le programme sont sélectionnés sur la base d'un indice de pauvreté établi à partir de données du recensement et de données d'une enquête de référence. Ce programme social de grande envergure a été mis en place de manière progressive. Les deux tiers environ des localités (soit 314 sur 495) ont été choisies de manière aléatoire pour bénéficier du programme au cours des deux premières années. Les 181 localités restantes ont constitué un groupe de comparaison avant d'intégrer le programme la troisième année.

Sur la base de l'assignation aléatoire, Schultz (2004) conclut à une augmentation moyenne du taux de scolarisation de 3,4 % chez les écoliers de la première à la huitième année, la hausse la plus importante (soit 14,8 %) étant constatée chez les filles ayant terminé la sixième année[a]. Cette forte croissance est probablement due au fait que le taux d'abandon scolaire tend à augmenter chez les filles au fur et à mesure qu'elles grandissent ; raison pour laquelle les filles reçoivent une allocation monétaire un peu plus importante pour les inciter à continuer à fréquenter l'école au-delà du primaire. Ces impacts à court terme sont ensuite extrapolés pour prédire l'impact à long terme du programme Progresa sur la scolarité et sur les revenus.

Source : Schultz, 2004.

a. Pour être précis, Schultz combine les méthodes d'assignation aléatoire et de double différence. Le chapitre 8 montre l'intérêt de combiner diverses méthodes d'évaluation d'impact.

9. Classe de CE2 dans le système scolaire français.

10. Classe de 3ème dans le système scolaire français.

Offre aléatoire : lorsque tout le monde n'adhère pas à son affectation

Lorsque nous avons évoqué l'assignation aléatoire ci-dessus, nous avons supposé que le responsable de programme avait toute latitude pour affecter les unités au groupe de traitement et au groupe de comparaison, les premières participant au programme et les secondes n'y participant pas. Autrement dit, les unités des deux groupes *adhéraient* pleinement à leur affectation. Ce type d'adhérence totale est cependant plus fréquent dans des conditions de laboratoire ou lors d'essais médicaux. Par exemple, le chercheur peut s'assurer, d'une part, que tous les sujets du groupe de traitement prennent bien leurs comprimés et, d'autre part, qu'aucun sujet du groupe de comparaison n'en prend[11].

Dans le cadre des programmes sociaux, l'adhérence totale aux critères de sélection (c'est-à-dire l'adhérence totale des unités à leur assignation au groupe de comparaison ou de traitement) est optimale, et tant les décideurs que les évaluateurs font au mieux pour se rapprocher au plus près de cet idéal. En pratique, pourtant, il n'est pas garanti que toutes les unités respectent pleinement leur affectation au groupe désigné, et ce malgré les meilleurs efforts des évaluateurs et des décideurs politiques. Par exemple, il ne suffit pas qu'un enseignant soit affecté au groupe de traitement et qu'une formation lui soit proposée pour qu'il se présente effectivement le jour du début de cette formation. De même, un enseignant du groupe de comparaison peut trouver un moyen de participer à une formation à laquelle il n'a pas été invité. Dans ces conditions, une comparaison directe des unités initialement affectées au groupe de traitement avec celles initialement assignées au groupe de comparaison donnera une *estimation de « l'intention de traiter » (IDT)*. En effet, la différence entre les deux groupes compare les unités à qui nous avions l'intention d'offrir un traitement (groupe de traitement) avec celles à qui nous n'avions pas l'intention d'offrir le traitement (groupe de comparaison). En tant que telle, l'estimation de « l'intention de traiter » constitue une mesure d'impact tout à fait pertinente, car, dans la plupart des cas, les décideurs politiques et les responsables de programme ne peuvent qu'offrir le programme à des bénéficiaires potentiels et non imposer à la population cible d'y participer.

Cependant, nous pouvons aussi chercher à connaître l'impact du programme sur ceux qui ont effectivement accepté d'y participer. Pour ce faire, il convient de prendre en compte le fait que certaines unités du groupe de traitement n'ont pas, dans les faits, été soumises au traitement et qu'inversement, certaines unités du groupe de comparaison y ont été soumises. En d'autres termes, nous souhaitons estimer l'impact du programme pour les unités auxquelles le programme a été offert et qui ont effectivement choisi d'y participer, autrement dit, *l'estimation du « traitement sur les traités » (TT)*.

Offre aléatoire d'un programme et participation effective

Imaginez que vous devez évaluer l'impact d'un programme de formation professionnelle sur les salaires. Le programme fait l'objet d'une assignation aléatoire au niveau

individuel, et le groupe de traitement se voit offrir la formation, contrairement au groupe de comparaison. Dans ce contexte, il y a trois types d'individus :

- *Ceux qui participent si on le leur offre.* Il s'agit des personnes qui adhèrent à leur affectation. Si elles sont affectées au groupe de traitement (à qui le programme est offert), elles participent au programme ; si, en revanche, elles sont affectées au groupe de comparaison (à qui le programme n'est pas offert), elles n'y participent pas.
- *Les « jamais ».* Il s'agit des personnes qui ne participent pas au programme même si elles sont assignées au groupe de traitement. Elles constituent les non adhérents dans le groupe de traitement.
- *Les « toujours ».* Il s'agit des personnes qui trouvent un moyen de bénéficier du programme même si elles sont affectées au groupe de comparaison. Elles constituent les non adhérents dans le groupe de comparaison.

Dans l'exemple du programme de formation professionnelle, le groupe des *jamais* peut être constitué de personnes non motivées qui, même si on leur a offert une formation, ne se présenteront pas. Au contraire, le groupe des *toujours* peut être constitué de personnes tellement motivées qu'elles trouveront un moyen de bénéficier du programme même si elles ont été initialement assignées au groupe de comparaison. Enfin, le groupe de *ceux qui participent si on le leur offre* comprend les personnes qui viendront à la formation si celle-ci leur est offerte (groupe de traitement), mais qui ne chercheront pas à participer si elles font partie du groupe de comparaison.

La figure 4.6 représente l'offre aléatoire du programme et la participation effective de ces trois groupes (*ceux qui participent si on le leur offre, les jamais et les toujours*). Supposons que la population totale est composée de 80 % de personnes qui *participent si on le leur offre*, de 10 % de *jamais* et de 10 % de *toujours*. Si l'échantillon d'évaluation est un échantillon aléatoire de la population, cet échantillon sera lui aussi composé approximativement de 80 % de personnes qui *participent si on le leur offre*, de 10 % de *jamais* et de 10 % de *toujours*. Si nous répartissons ensuite les unités de l'échantillon d'évaluation entre groupe de traitement et groupe de comparaison, ces mêmes proportions se maintiennent (80 % qui *participent si on le leur offre*, 10 % de *jamais* et 10 % de *toujours*). Dans le groupe à qui le programme est offert, deux groupes participent au programme (ceux qui *participent si on le leur propose* et les *toujours*), alors que les *jamais* restent à l'écart. Dans le groupe à qui le programme n'est pas offert, seuls les *toujours* intègrent le programme, mais pas *ceux qui ne participent que si on le leur propose* ni les *jamais*.

Estimation d'impact pour l'offre aléatoire

Maintenant que nous avons établi la différence entre l'offre d'un programme et la participation effective au programme, nous allons nous intéresser à une technique qui peut être utilisée pour estimer l'impact du traitement sur les traités, autrement

Figure 4.6 Offre aléatoire d'un programme

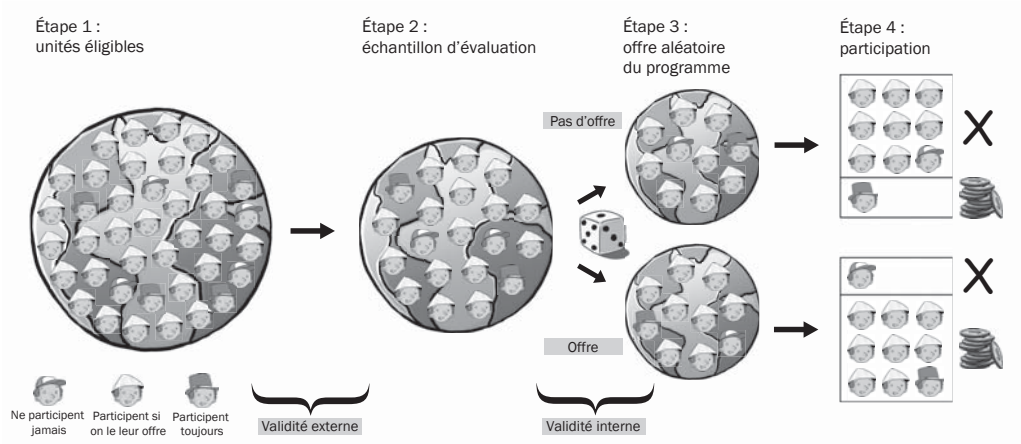
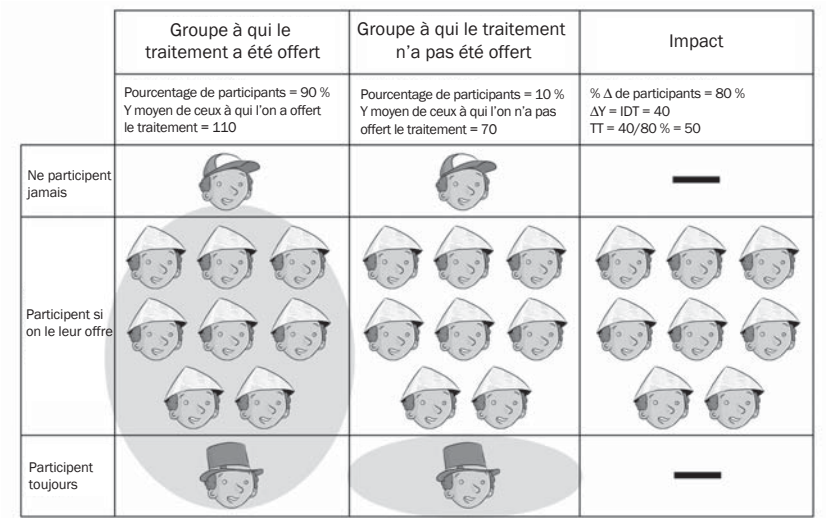


Figure 4.7 Estimation de l'impact du traitement sur les traités en cas d'offre aléatoire



Remarque : l'IDT, estimation de « l'intention de traiter », est obtenue en comparant les résultats du groupe auquel le traitement a été offert à ceux du groupe auquel le traitement n'a pas été offert (indépendamment de la participation effective). Le TT correspond à l'estimation du « traitement sur les traités » c'est-à-dire à l'estimation de l'impact sur ceux à qui le programme a été offert et qui y ont effectivement participé. Les personnages sur fond grisé sont ceux qui participent effectivement au programme.

dit l'impact d'un programme sur ceux à qui le programme a été offert *et* qui y ont effectivement participé. Cette estimation s'effectue en deux étapes, présentées dans la figure 4.7[12].

En premier lieu, nous procédons à l'estimation de l'impact de l'intention de traiter. Souvenez-vous qu'il s'agit de la différence entre l'indicateur de résultat Y du groupe auquel on a offert le traitement et le même indicateur pour le groupe auquel on n'a pas offert le traitement. Par exemple, si le revenu moyen (Y) est de 110 dollars pour le groupe de traitement et de 70 dollars pour le groupe de comparaison, l'estimation d'impact de l'intention de traiter (IDT) est alors de 40 dollars.

En second lieu, nous devons déduire l'estimation du traitement sur les traités (TT) à partir de l'estimation de l'intention de traiter (IDT). Pour ce faire, nous devons déterminer d'où vient la différence de 40 dollars. Procédons par élimination. Nous savons que la différence ne peut pas être attribuée à une quelconque différence entre les *jamais* du groupe de traitement (à qui le programme a été offert) et du groupe de comparaison (à qui le programme n'a pas été offert). Comme les *jamais* ne sont par définition pas concernés par le programme, il n'y a pour eux pas de différence qu'ils soient dans le groupe de traitement ou dans le groupe de comparaison. Nous savons aussi que la différence de 40 dollars ne peut pas être due à des différences entre les *toujours* des groupes de traitement et de comparaison, puisqu'ils participent dans les deux cas. Pour eux aussi, peu importe qu'ils fassent partie du groupe de traitement ou du groupe de comparaison. Par conséquent, la différence de résultat constatée entre les deux groupes ne peut provenir que des effets du programme sur le seul groupe dont le comportement est modifié par son affectation au groupe de traitement ou au groupe de comparaison, à savoir *ceux qui participent si on le leur offre*. Si nous arrivons à identifier *ceux qui participent si on le leur offre*, il sera facile d'estimer l'impact du programme sur ces unités.

Dans les faits, bien que nous sachions que ces trois types d'individus existent dans la population, nous ne pouvons pas séparer les personnes selon leur appartenance au groupe de *ceux qui participent si on le leur offre*, des *jamais* ou des *toujours*. Dans le groupe qui s'est vu offrir le traitement, nous pouvons repérer les *jamais* (car ils ne participent pas), mais il est impossible de faire la distinction entre les *toujours* et *ceux qui participent si on le leur offre* (car tous deux participent ensemble). Inversement, dans le groupe auquel le traitement n'a pas été offert, nous pouvons isoler les *toujours* (car ils ont intégré le programme), mais on ne peut faire la distinction entre les *jamais* et *ceux qui participent si on le leur propose*.

Toutefois, en sachant que 90 % des unités du groupe auquel le traitement a été offert y participent effectivement, nous pouvons déduire que 10 % des unités dans la population sont des *jamais* (soit la partie des personnes à qui le programme a été offert, mais qui n'y participent pas). De même, en constatant que 10 % des unités auxquelles le traitement n'a pas été offert y participent quand même, il est possible de conclure que ces 10 % représentent des *toujours* (soit la partie des individus du groupe à qui le programme n'a pas été offert, mais qui l'ont tout de même intégré). Il reste alors 80 % des unités dans le groupe de *ceux qui participent si on le leur offre*.

Nous savons que la totalité de l'impact de 40 dollars est due à la différence de participation des 80 % d'unités de notre échantillon, *ceux qui participent si on le leur offre*. Si 80 % des unités sont à l'origine de l'impact moyen de 40 dollars constaté pour l'ensemble du groupe à qui le traitement a été offert, l'impact sur ces 80 % de personnes qui *participent si on le leur offre* est de 40/0,8, soit 50 dollars. Autrement dit, l'impact du programme sur *ceux qui participent si on le leur offre* est de 50 dollars, mais lorsque cet impact est considéré pour l'ensemble du groupe auquel le traitement a été offert, il se dilue de 20 % à cause des unités qui n'ont pas adhéré à l'assignation aléatoire initiale.

L'un des problèmes fondamentaux avec l'auto-sélection des individus dans les programmes est qu'il n'est pas toujours possible de savoir pourquoi certaines personnes choisissent de participer et d'autres non. Lorsque nous procédons à une sélection aléatoire des unités qui vont participer au programme, mais que la participation effective dépend de la volonté de chacun et qu'il existe un moyen pour les unités assignées au groupe de comparaison de bénéficier tout de même du programme, nous sommes confrontés à un problème similaire : nous ne serons pas toujours en mesure de comprendre le processus qui conduit certaines personnes à ne *jamais participer*, à *toujours participer* ou à *participer si on le leur offre* comme dans l'exemple ci-dessus. Toutefois, pour autant que ceux qui n'adhèrent pas à leur affectation ne soient pas trop nombreux, l'assignation aléatoire initiale demeure un outil efficace d'estimation d'impact. L'inconvénient du manque d'adhérence totale des individus est que l'estimation d'impact ne pourra plus être considérée comme valide pour l'ensemble de la population. Cette estimation ne sera valable que pour un sous-groupe spécifique de la population cible, à savoir celui des individus qui *participent si on le leur offre*.

L'offre aléatoire présente deux caractéristiques importantes qui permettent d'estimer l'impact, même à défaut d'une adhérence totale (voir encadré 4.2)[13].

1. Elle peut servir pour prédire la participation effective au programme si la plupart des individus se comportent comme *ceux qui participent si on le leur offre*, c'est-à-dire qui intègrent le programme si celui-ci leur est offert, mais qui ne le font pas dans le cas contraire.

2. Les deux groupes (celui à qui le traitement est offert et celui à qui il n'est pas offert) étant constitués à partir d'un processus de sélection aléatoire, les caractéristiques des individus des deux groupes ne sont corrélées avec aucun autre élément, par exemple les capacités ou la motivation, qui aurait aussi pu affecter le résultat (Y).

Promotion aléatoire ou modèle d'encouragement

Dans la section précédente, nous avons vu comment estimer l'impact d'un programme dans le cas d'une assignation aléatoire du traitement, même si les affectations initiales aux groupes de comparaison et au groupe de traitement ne sont pas totalement respectées. Nous allons maintenant examiner une approche très similaire qui peut être utilisée pour évaluer les programmes à éligibilité universelle, à participation volontaire, ou pour lesquels il n'est pas possible de déterminer qui participe et qui ne participe pas.

Encadré 4.2 : Offre aléatoire de bons d'éducation en Colombie

En Colombie, le Programme d'extension de la couverture de l'éducation secondaire (Programa de Ampliación de Cobertura de la Educación Secundaria [PACES]) a permis à plus de 125 000 étudiants de bénéficier de bons leur permettant de couvrir un peu plus de la moitié du coût de leur scolarisation dans une école secondaire privée. Le budget du programme PACES étant limité, ces bons ont été attribués par tirage au sort. Angrist et al. (2002) profitent de cette assignation aléatoire du traitement pour déterminer l'impact du programme de distribution de bons sur des indicateurs de résultats sociaux et éducatifs.

Ils aboutissent à la conclusion que les étudiants tirés au sort sont dix points plus susceptibles de terminer la 8ème année[14] et affichent une moyenne aux examens standardisés de 0,2 écart-type supérieur trois ans après le tirage au sort. Ils découvrent également que les effets du programme éducatif sont plus marqués pour les filles que pour les garçons. Les chercheurs examinent ensuite l'impact du programme sur plusieurs résultats au-delà de l'éducation et trouvent que les personnes tirées au sort sont moins susceptibles d'être mariées et travaillent environ 1,2 heure de moins par semaine.

Dans le contexte de cette étude, l'adhérence à l'assignation aléatoire n'est pas totale puisque seuls 90 % environ des personnes tirées au sort ont utilisé les bons ou une autre forme de bourse scolaire, et que 24 % des personnes non tirées au sort ont tout de même reçu une bourse scolaire. Angrist et ses collaborateurs utilisent donc également l'intention de traiter (en l'occurrence si l'étudiant a été tiré au sort ou non) en tant que variable instrumentale pour déterminer le traitement sur les traités, soit la réception effective d'une bourse scolaire. Finalement, les chercheurs effectuent également une analyse coût-bénéfice pour mieux comprendre l'impact du programme de bons d'éducation sur les dépenses à la fois des ménages et de l'État. Ils concluent que le coût social total du programme est limité, mais largement compensé par les retours espérés pour les participants et leur famille. Ceci suggère que des programmes qui, comme le PACES, sont axés sur la demande peuvent constituer un moyen efficace et rentable d'améliorer l'accès à l'éducation.

Source : Angrist et al. 2002.

Les gouvernements mettent souvent en œuvre des programmes pour lesquels il est difficile d'exclure des participants potentiels ou de les forcer à participer. De nombreux programmes permettent aux participants potentiels de choisir de participer ou non et ne peuvent, par conséquent, exclure les participants potentiels désirant y participer. Par ailleurs, certains programmes sont dotés d'un budget suffisamment important pour couvrir immédiatement l'ensemble de la population éligible. Dans ce cas, affecter certains participants de manière aléatoire à un groupe de traitement ou à un groupe de comparaison, et en exclure certains aux fins de l'évaluation ne serait pas éthiquement acceptable. Nous avons donc besoin d'une autre méthode pour évaluer l'impact de ce genre de programme (c'est-à-dire les programmes à participation volontaire ou à éligibilité universelle).

Les programmes à participation volontaire laissent généralement le choix aux personnes intéressées de s'y inscrire et d'y participer. Revenons à l'exemple du programme de formation professionnelle évoqué auparavant, mais imaginons cette

fois-ci qu'une assignation aléatoire n'est pas possible et que toute personne souhaitant bénéficier du programme peut s'y inscrire. Comme précédemment, il est fort probable d'avoir à faire à trois types d'individus : les adhérents, des individus qui ne participent *jamais* et des individus qui participent *toujours*. Comme dans le cas précédent, les *« toujours »* intégreront le programme dans tous les cas alors que les *« jamais »* ne s'y joindront en aucun cas. Mais qu'en est-il des adhérents ? Dans le cas présent, toute personne souhaitant participer au programme est libre de le faire. Qu'en est-il des personnes qui pourraient être très intéressées par le programme, mais qui, pour diverses raisons, n'auront, par exemple, pas suffisamment d'information ou de motivation pour y participer ? Dans ces conditions, les adhérents seront de *ceux qui participent en cas de promotion* : il s'agit d'un groupe d'individus qui participent au programme s'il existe des incitations supplémentaires (c.-à-d. une forme de promotion) les amenant à participer. À défaut de ces incitations supplémentaires, *ceux qui participent en cas de promotion* n'intégreront pas le programme.

Revenons à l'exemple de la formation professionnelle. Si l'agence qui organise la formation dispose des fonds et des capacités nécessaires pour dispenser la formation à toute personne intéressée, le programme pourra alors être ouvert à toute personne au chômage qui désire y participer. Il est cependant peu probable que toutes les personnes au chômage souhaitent se former ou même qu'elles soient toutes au courant de l'existence du programme. Certains chômeurs peuvent être réticents à participer au programme parce qu'ils ne disposent pas de suffisamment d'informations sur le contenu de la formation et qu'ils ne parviennent pas à trouver d'informations supplémentaires. Supposons maintenant que l'agence qui dispense cette formation engage une assistante communautaire pour faire une promotion de ce programme de formation professionnelle. Munie d'une liste des chômeurs, elle se rend au domicile des personnes concernées, leur décrit le programme de formation et leur propose de s'y inscrire de suite. Bien évidemment, elle ne peut forcer personne à y participer. Par ailleurs, certains chômeurs qui n'auront pas reçu la visite de l'assistante pourront aussi s'inscrire à la formation, mais ils devront s'adresser directement à l'institut de formation. Nous sommes désormais face à deux groupes de chômeurs : ceux qui ont reçu la visite de l'assistante et ceux qui ne l'ont pas reçue. Si l'effort de promotion du programme auprès de la population a porté ses fruits, le taux de participation des chômeurs ayant reçu la visite de l'assistante devrait être supérieur à celui des chômeurs n'ayant pas été contactés par l'assistante.

Comment pourrions-nous évaluer l'impact du programme de formation ? Comme nous le savons, il ne suffit pas de comparer les chômeurs ayant suivi la formation à ceux qui ne l'ont pas suivie, car les chômeurs ayant décidé de s'inscrire présentent probablement des caractéristiques, tant observables que non observables, très différentes des caractéristiques de ceux qui ne participent pas au programme: ils peuvent avoir un niveau d'éducation plus élevé (caractéristique facilement observable) et ils peuvent être plus motivés par l'idée de trouver un emploi (caractéristique difficile à observer et à mesurer).

Nous disposons néanmoins d'une variable supplémentaire à exploiter pour trouver un groupe de comparaison valide. Examinons tout d'abord s'il est possible de comparer le groupe ayant reçu la visite de l'assistante avec celui qui ne l'a pas reçue.

Les deux groupes comprennent des personnes très motivées (les *« toujours »*) qui intégreront la formation qu'ils aient ou non reçu la visite de l'assistante. De même, dans les deux groupes, nous retrouverons des personnes non motivées (les *« jamais »*) qui ne participeront pas au programme, quels que soient les efforts de l'assistante. Enfin, certaines personnes (*ceux qui participent en cas de promotion*) rejoindront la formation si l'assistante leur rend visite, mais pas dans le cas contraire.

Si l'assistante a sélectionné les personnes auxquelles elle rend visite aléatoirement à partir de sa liste de chômeurs, nous pourrons avoir recours à la méthode du traitement sur les traités évoquée ci-dessus. La seule différence est qu'il s'agit ici non plus d'une *offre* aléatoire, mais d'une promotion aléatoire du programme. À partir du moment où il existe des personnes qui *ne participent qu'en cas de promotion* (c'est-à-dire dont la participation n'est assurée que si on va les « chercher »), il y aura une variation entre le groupe *avec* promotion et le groupe *sans* promotion qui nous permettra d'estimer l'impact de la formation sur *ceux qui y participent en cas de promotion*. Au lieu d'adhérer à l'offre de traitement, *ceux qui participent en cas de promotion* adhèrent à la promotion du programme.

D'un côté, la stratégie de promotion doit être efficace et entraîner une nette augmentation des inscriptions de *ceux qui participent en cas de promotion*. D'un autre côté, nous ne souhaitons pas que les activités de promotion soient efficaces au point d'influencer le résultat. Par exemple, si les assistantes chargées de la promotion proposent des sommes d'argent importantes aux chômeurs pour les inciter à s'inscrire, il sera difficile d'établir plus tard si les variations de revenus constatées sont dues à la formation, à la promotion du programme ou aux incitations proposées.

La promotion aléatoire est une stratégie qui permet de générer l'équivalent d'un groupe de comparaison aux fins de l'évaluation. Elle peut être utilisée lorsqu'il est possible d'organiser une campagne de promotion visant un échantillon aléatoire de la population cible. Les lecteurs ayant quelques connaissances en économétrie reconnaîtront la terminologie introduite dans la section précédente : la promotion aléatoire est une variable instrumentale permettant de créer une variation entre les unités et d'exploiter cette variation pour créer un groupe de comparaison valide.

Vous avez dit « promotion » ?

La promotion aléatoire vise à accroître la participation des individus d'un sous-échantillon de population à un programme volontaire. Elle peut prendre plusieurs formes. Il peut par exemple s'agir d'une campagne d'information à l'attention des personnes qui ne se sont pas inscrites, car elles ne connaissaient pas ou ne comprenaient pas bien le contenu du programme. La promotion peut aussi comprendre des incitations comme l'offre de petits cadeaux ou prix, ou encore la mise à disposition de moyens de transport.

Plusieurs conditions doivent être remplies pour que la méthode de promotion aléatoire permette une évaluation d'impact valide.

1. Les groupes recevant la promotion et ceux ne la recevant pas doivent être comparables. Ils doivent présenter des caractéristiques similaires. Ceci est assuré grâce à une assignation aléatoire des activités de promotion aux unités de l'échantillon d'évaluation.
2. La campagne de promotion doit augmenter la participation au programme des individus qui la reçoivent en comparaison aux individus qui ne la reçoivent pas. Pour s'assurer que c'est effectivement le cas, il suffit de vérifier que le taux de participation est plus élevé dans le groupe ayant bénéficié de la promotion que dans l'autre.
3. Il est important que les activités de promotion n'aient pas un impact direct sur les résultats ; il faut en effet pouvoir attribuer lesdits résultats au programme et non aux activités de promotion.

Concept clé :

La promotion aléatoire est une méthode similaire à l'offre aléatoire. Toutefois, au lieu de sélectionner de manière aléatoire les unités auxquelles le traitement sera offert, nous sélectionnions ici, toujours aléatoirement, les unités qui recevront une promotion du programme. Le programme reste alors ouvert à toutes les unités.

Le processus de promotion aléatoire

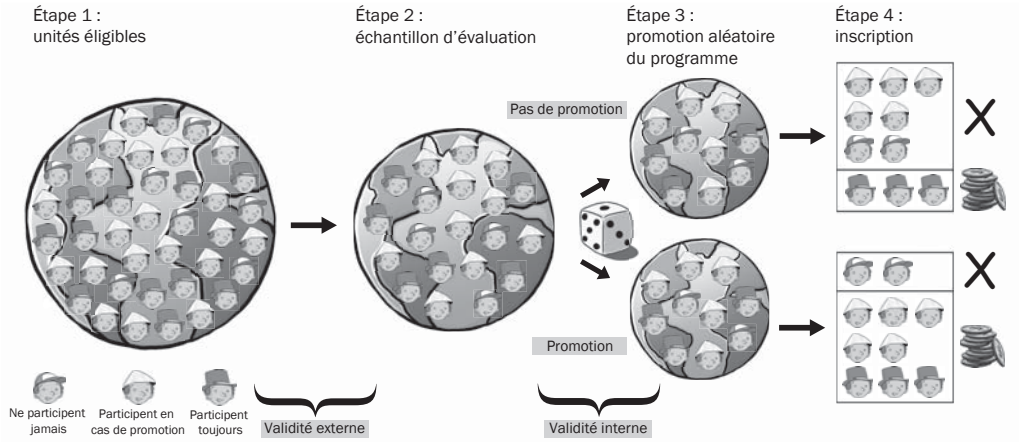
Le processus de promotion aléatoire est présenté à la figure 4.8. Comme pour les méthodes précédentes, nous partons de la population des unités éligibles au programme. Contrairement à la méthode de l'assignation aléatoire, nous ne pouvons plus sélectionner de manière aléatoire qui participera au traitement et qui n'y participera pas, la participation au programme étant dorénavant entièrement volontaire. Nous savons toutefois qu'il y aura trois types d'unités au sein de la population des unités éligibles :

- *Les toujours* — les personnes qui participeront au programme dans tous les cas.
- *Ceux qui participent en cas de promotion* — les personnes qui ne s'inscriront au programme que si elles reçoivent la promotion du programme.
- *Les jamais* — les personnes qui n'intégreront pas le programme, promotion ou pas.

Soulignons à nouveau que l'appartenance des unités à l'un de ces trois groupes est une caractéristique intrinsèque que l'évaluateur ne peut observer, car elle est liée à des facteurs comme la motivation ou l'intelligence de chacun.

Une fois la population éligible déterminée, l'étape suivante consiste à sélectionner de manière aléatoire un échantillon d'évaluation à partir de la population. Les unités de l'échantillon sont celles pour lesquelles des données seront collectées. Dans certains cas, toute la population pourra être incluse dans l'échantillon d'évaluation, par exemple si nous disposons de données sur l'ensemble de la population des unités éligibles.

Figure 4.8 Promotion aléatoire



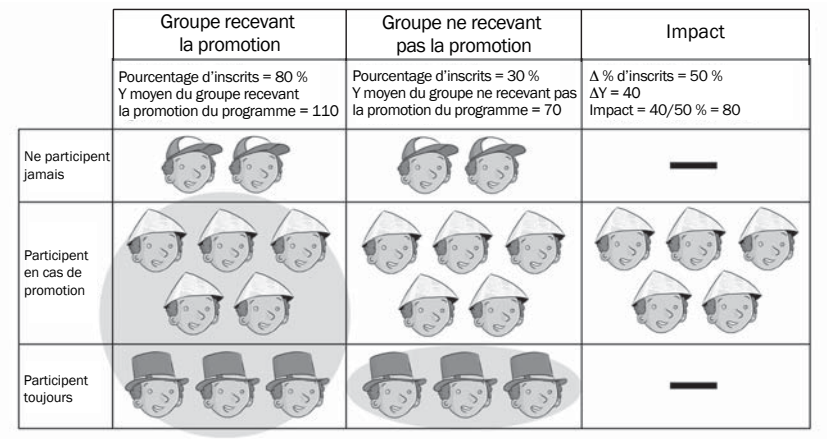
Une fois l'échantillon d'évaluation déterminé, la promotion aléatoire consiste à répartir de façon aléatoire les unités de cet échantillon entre le groupe qui recevra la promotion et le groupe qui ne la recevra pas. Comme la sélection est effectuée de manière aléatoire, les membres des deux groupes présenteront les mêmes caractéristiques que ceux de l'échantillon d'évaluation, et ces caractéristiques seront également équivalentes à celles de l'ensemble de la population des unités éligibles. Le groupe bénéficiant de la promotion et le groupe n'en bénéficiant pas auront donc des caractéristiques similaires.

Après la campagne de promotion, nous pouvons examiner les taux de participation de chaque groupe. Au sein du groupe qui n'a pas reçu de promotion, seuls les *toujours* intégreront le programme. Nous saurons alors qui sont les *toujours* dans le groupe qui n'a pas reçu de promotion, mais dans ce même groupe nous ne pourrons pas distinguer les *jamais* de *ceux qui participent en cas de promotion*. À l'inverse, dans le groupe qui a reçu la promotion, *ceux qui participent en cas de promotion* et les *toujours* intégreront le programme, tandis que les *jamais* resteront à l'écart. Dans ce groupe, nous pourrons donc identifier les *jamais*, mais il sera impossible de faire la distinction entre *ceux qui participent en cas de promotion* et les *toujours*.

Estimation d'impact pour la promotion aléatoire

L'estimation de l'impact d'un programme faisant l'objet d'une promotion aléatoire est un cas particulier de la méthode de traitement des traités (figure 4.9). Imaginons que le taux de participation soit de 30 % dans le groupe n'ayant pas reçu la campagne de promotion (trois *toujours*), mais qu'il atteigne 80 % dans le groupe ciblé par la campagne de promotion (trois *toujours* et cinq individus *qui participent en cas de promotion*). Supposons que le résultat moyen soit de 70 pour les personnes du groupe non soumis à une promotion (dix individus) et de 110 pour ceux du groupe touché par la promotion (dix individus). Dans ce cas, quel sera l'impact du programme ?

Figure 4.9 Estimation d'impact en cas de promotion aléatoire



Remarque : les personnages sur fond grisé sont ceux qui participent au programme.

Premièrement, nous connaissons la différence entre le groupe qui a reçu la promotion du programme et celui qui ne l'a pas reçu : elle est de 40. Nous savons aussi que cette différence ne peut pas être due aux *jamais*, car, dans tous les cas, ils ne participeront pas au programme. Cette différence ne peut pas non plus être attribuée aux *toujours* parce qu'ils participent au programme quel que soit le groupe auquel ils appartiennent initialement.

La deuxième étape consiste à déterminer l'impact du programme sur *ceux qui participent en cas de promotion*. Nous savons que tout l'effet moyen (de 40) peut être attribué à ceux qui *participent en cas de promotion*, un groupe qui représente la moitié de la population. Pour évaluer l'impact moyen du programme sur une personne adhérant aux règles d'affectation, nous divisons alors 40 par le pourcentage de *ceux qui participent en cas de promotion* dans la population. Nous ne pouvons certes pas identifier directement ce dernier groupe, mais nous pouvons évaluer sa part dans la population : elle correspond à la différence entre les taux de participation du groupe auprès duquel la promotion a été réalisée et du groupe pour lequel ça n'a pas été le cas (50 % ou 0,5). L'impact moyen sur une personne adhérant aux règles d'affectation s'établit donc à 40/0,5 = 80.

La promotion étant effectuée de manière aléatoire auprès des individus, le groupe qui a bénéficié de cette promotion et le groupe qui n'en a pas bénéficié présenteront des caractéristiques moyennes identiques. Dès lors, les différences entre les résultats moyens des deux groupes peuvent être attribuées au fait que dans le groupe recevant la promotion, *ceux qui participent en cas de promotion* ont effectivement participé au programme alors qu'ils ne l'ont pas fait dans le groupe n'ayant pas reçu la campagne de promotion[15].

Impact du Programme de subvention de l'assurance maladie (PSAM) selon la méthode de la promotion aléatoire

Nous allons maintenant appliquer la méthode de la promotion aléatoire pour évaluer l'impact du PSAM. Supposons que le ministère de la Santé décide que les subventions soient distribuées immédiatement à tout ménage souhaitant participer au PSAM. Vous savez toutefois que la couverture nationale ne peut être atteinte que graduellement. Vous vous mettez d'accord avec le ministère de la Santé pour accélérer la participation par le biais d'une campagne de promotion dans un groupe de villages choisis aléatoirement. Vous mettez en place une grande campagne de promotion (communication et marketing social) ciblant ce groupe de villages et visant à sensibiliser les habitants au PSAM. Après deux années d'efforts promotionnels et de mise en œuvre du programme, il apparaît que 49,2 % des ménages des villages ayant bénéficié de la campagne de promotion ont rejoint le programme, contre 8,4 % seulement dans les villages qui n'ont pas été touchés par la campagne (tableau 4.4).

La sélection des villages auprès desquels a eu lieu la campagne de promotion ayant été réalisée de manière aléatoire, les caractéristiques moyennes des deux groupes auraient été les mêmes en l'absence du programme.

Cette hypothèse peut être vérifiée en comparant les dépenses de santé (ainsi que d'autres caractéristiques) des deux groupes au moment l'enquête de base. Deux années après la mise en œuvre du programme, les dépenses de santé moyennes dans les villages ayant été soumis à la campagne de promotion sont de 14,9 dollars contre 18,8 dollars dans les zones non couvertes par cette campagne (soit – 3,9 dollars de différence). Toutefois, comme la seule différence entre les villages touchés par la campagne de promotion et les autres est un taux de participation plus élevé dans les premiers (grâce aux efforts de promotion), la différence de 3,9 dollars dans les dépenses de santé peut être attribuée aux 40,4 % de ménages des villages recevant la promotion qui se sont inscrits grâce à elle. Nous devons donc ajuster la différence dans les dépenses de santé pour évaluer l'impact du programme sur *ceux qui partici-*

Tableau 4.4 Cas 4— Impact du PSAM selon la méthode de promotion aléatoire (comparaison de moyennes)

	Villages ayant reçu la campagne de promotion	Villages n'ayant pas reçu la campagne de promotion	Différence	Stat. de *t*
Dépenses de santé des ménages observées lors de l'enquête de base	17,1	17,2	–0,1	–0,47
Dépenses de santé des ménages observées lors de l'enquête de suivi	14,9	18,8	–3,9	–18,3
Participation au PSAM	49,2%	8,4%	40,4%	

** Seuil de signification de 1 %.

Tableau 4.5 Cas 4— Impact du PSAM selon la méthode de promotion aléatoire (analyse de régression)

	Régression linéaire	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–9,4** (0,51)	–9,7** (0,45)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

pent en cas de promotion. Pour ce faire, nous divisons la différence observée entre les groupes par le pourcentage de *ceux qui participent en cas de promotion* : – 3,9/0,404 = – 9,65 $. Votre collègue, qui a suivi des cours d'économétrie, calcule ensuite l'impact du programme par la méthode des moindres carrés en deux étapes et aboutit aux résultats présentés dans le tableau 4.5. L'impact ainsi estimé est valable pour les ménages ayant participé au programme parce qu'ils y ont été incités, mais qui n'y aurait pas participé sans promotion, autrement dit pour *ceux qui participent en cas de promotion*. Extrapoler ce résultat à l'ensemble de la population suppose que tous les autres ménages se seraient comportés de la même manière s'ils avaient intégré le programme.

QUESTION 4

A. Quelles sont les hypothèses de base qui sous-tendent le résultat du cas 4 ?
B. Au vu de ces résultats pour le cas 4, le PSAM doit-il être élargi à l'échelle nationale ?

La promotion aléatoire en pratique

La méthode de la promotion aléatoire a été utilisée dans plusieurs contextes. Gertler, Martinez et Vivo (2008) y ont eu recours pour évaluer un programme d'assurance de santé maternelle et infantile en Argentine. Après la crise économique de 2001, l'État argentin a constaté que les indicateurs de santé de la population se dégradaient avec, notamment une augmentation de la mortalité infantile. Il a décidé d'introduire un système d'assurance national pour les mères et les enfants qui devaient s'étendre à l'ensemble du pays en un an. Avant cela, les autorités ont souhaité évaluer l'impact du programme pour s'assurer qu'il entraînait bien une amélioration de la santé de la population. Comment trouver un groupe de comparaison si chaque mère et chaque enfant du pays sont éligibles pour participer au système d'assurance s'ils le souhaitaient ? Les données provenant des premières provinces ayant mis en œuvre l'intervention ont montré que seulement 40 % à 50 % des ménages s'étaient effectivement inscrits au programme. Les autorités ont alors lancé une vaste campagne de promotion visant à informer les populations sur le programme. Cette campagne n'a toutefois touché que certains villages, sélectionnés sur une base aléatoire, et non l'ensemble du pays.

Il existe d'autres exemples comme l'aide apportée par des organisations non gouvernementales dans le cadre de l'évaluation de la gestion scolaire communautaire au Népal ou le Fonds d'investissement social en Bolivie (décrit dans l'encadré 4.3).

Limites de la méthode de la promotion aléatoire

La promotion aléatoire est une stratégie utile pour évaluer l'impact des programmes à participation volontaire et à éligibilité universelle, notamment parce qu'elle n'exige d'exclure aucune des unités éligibles. Cette approche présente néanmoins quelques limites en comparaison à l'assignation aléatoire du traitement.

Premièrement, la stratégie de promotion doit porter ses fruits. Si la campagne de promotion n'entraîne pas d'augmentation de la participation, aucune différence ne ressortira entre le groupe recevant la promotion et celui ne la recevant pas ; aucune comparaison ne sera alors possible. Un suivi rapproché de la campagne promotionnelle est donc primordial pour en assurer l'efficacité. Le point positif est que la conception de la campagne de promotion peut permettre aux responsables du programme de réfléchir à la manière dont ils peuvent encourager la participation.

Encadré 4.3 : Promotion des investissements dans les infrastructures d'éducation en Bolivie

En 1991, la Bolivie met en place un Fonds d'investissement social (FIS) visant à fournir des financements aux communautés rurales pour qu'elles réalisent des investissements de petite envergure dans des infrastructures d'éducation, de santé et d'eau. En parallèle, la Banque mondiale, qui contribue au financement du FIS, met en place une évaluation d'impact prospective dès la conception du programme.

Dans le cadre de l'évaluation d'impact du volet portant sur l'éducation, une sélection aléatoire est effectuée au sein des communautés de la région du Chaco pour déterminer celles qui bénéficient d'une promotion du FIS, à travers des visites et des encouragements supplémentaires pour les inciter à y adhérer. Le programme est ouvert à toutes les communautés éligibles de la région, sous réserve qu'elles fassent la démarche de présenter une demande pour la mise en œuvre d'un projet précis. Toutes les communautés ne participent pas au programme, mais les demandes sont supérieures chez les communautés ayant fait l'objet de la campagne de promotion.

Newman et al. (2002) utilisent la promotion aléatoire comme variable instrumentale. Ils concluent que les investissements dans l'éducation ont permis d'améliorer les indicateurs de qualité des infrastructures scolaires tels que l'électricité, les installations sanitaires, le nombre de manuels scolaires par élève et le nombre d'enseignants par élève. L'impact sur les résultats liés à l'éducation se révèle en revanche limité, à l'exception d'une baisse de 2,5 % du taux d'abandon scolaire. Forts de ces conclusions, le ministère de l'Éducation et le FIS réorientent les efforts et les ressources sur les aspects purement « éducatifs », ne finançant les améliorations d'infrastructures matérielles que dans le cadre d'interventions intégrées.

Source : Newman et al. 2002.

Deuxièmement, la méthode ne permet d'estimer l'impact d'un programme que pour un sous-groupe de la population des unités éligibles. Plus précisément, l'impact moyen du programme est calculé pour le groupe de personnes qui ont participé au programme uniquement parce qu'elles y ont été encouragées. Le problème est que les personnes composant ce groupe peuvent présenter des caractéristiques différentes de celles des individus qui participent toujours ou ne participent jamais. Aussi, l'impact moyen du traitement pour l'ensemble de la population peut être différent de l'impact moyen estimé pour les personnes qui ont participé parce qu'elles y ont été incitées.

Notes

1. L'assignation aléatoire du traitement est parfois appelée « essai contrôle randomisé », « évaluation aléatoire », « évaluation expérimentale » ou encore « expérimentation sociale ».
2. L'assignation aléatoire ne signifie pas qu'il y a forcément une chance sur deux d'être tiré au sort. En fait, la plupart des évaluations par assignation aléatoire donnent à chaque unité éligible une probabilité d'être sélectionnée déterminée de manière à ce que le nombre de gagnants (qui recevront le traitement) soit égal au nombre total de places offertes. Par exemple, si le programme dispose de suffisamment de fonds pour servir 1 000 communautés sur une population totale de 10 000, chaque communauté aura une chance sur dix d'être sélectionnée pour recevoir le traitement. La puissance statistique (concept évoqué en détail au chapitre 11) est optimisée lorsque l'échantillon d'évaluation est divisé à parts égales entre le groupe de traitement et le groupe de comparaison. Par exemple, pour un échantillon total comprenant 2 000 communautés, la puissance statistique sera optimisée en constituant un groupe de traitement de 1 000 communautés et un groupe de comparaison également de 1 000 communautés plutôt qu'en se fondant sur un simple échantillon aléatoire correspondant à 20 % des 10 000 communautés éligibles de départ (ceci donnerait un échantillon d'évaluation d'environ 200 communautés de traitement et 1 800 communautés de comparaison).
3. Par exemple, les programmes de logements subventionnés ont souvent recours aux tirages au sort pour sélectionner les bénéficiaires.
4. Cette propriété découle de la loi des grands nombres.
5. Un échantillon d'évaluation peut être stratifié par type d'individus et subdivisé en grappes d'unités. La taille de l'échantillon est fonction du type d'échantillonnage aléatoire utilisé (voir partie 3).
6. La plupart des logiciels permettent d'établir un « nombre source » (« seed number » en anglais) afin que les résultats de l'assignation aléatoire soient transparents et puissent être répétés.
7. Nous examinerons des concepts comme les effets de diffusion et de contamination de manière plus détaillée au chapitre 8.
8. Pour des raisons statistiques, il n'est pas nécessaire que toutes les caractéristiques observées soient similaires dans le groupe de traitement et dans le groupe de comparaison pour que la sélection aléatoire soit efficace. La règle d'or en

matière d'efficacité est que 95 % environ des caractéristiques observées soient similaires. Par « similaire », on entend que l'on ne peut rejeter l'hypothèse nulle selon laquelle les moyennes sont différentes entre les deux groupes compte tenu d'un intervalle de confiance de 95 %. Même lorsque les caractéristiques des deux groupes sont complètement égales, on peut s'attendre à ce que 5 % environ des caractéristiques présentent une différence statistiquement significative.

11. À noter que dans le domaine médical, les patients du groupe de comparaison reçoivent généralement un placebo, par exemple un comprimé en sucre sans effet sur les résultats. Ceci vise à tenir compte de « l'effet placebo », à savoir les changements éventuels de comportement et de résultats liés à la prise d'un traitement même si le traitement en soi n'a pas d'effet.
12. Ces deux étapes correspondent à la technique économétrique des moindres carrés en deux étapes qui permet d'obtenir l'estimation moyenne locale de l'effet du traitement (« local average treatment effect », ou LATE en anglais).
13. Les lecteurs ayant des connaissances en économétrie auront reconnu le concept : en statistiques, l'offre aléatoire du programme est utilisée comme variable instrumentale pour la participation effective. Les deux caractéristiques citées correspondent exactement à ce qui serait exigé d'une bonne variable instrumentale :
    - La variable instrumentale doit être corrélée à la participation au programme.
    - La variable instrumentale peut ne pas être corrélée au résultat (Y) (sauf par le biais de la participation au programme) ou aux variables non observables.
14. Classe de 4ème dans le système scolaire français.
15. Les lecteurs ayant des connaissances en économétrie comprendront que l'impact est estimé en utilisant « l'assignation aléatoire au groupe recevant ou ne recevant pas la promotion » comme variable instrumentale pour la participation effective au programme.

Références

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King et Michael Kremer. 2002. « Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. » *American Economic Review* 92 (5): 1535–58.

Gertler, Paul, Sebastian Martinez et Sigrid Vivo. 2008. « Child-Mother Provincial Investment Project *Plan Nacer*. » University of California Berkeley et Banque mondiale, Washington, DC.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa et Jose Luis Evia. 2002. « An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund. » *Étude économique de la Banque mondiale* 16 (2) : 241–74.

Schultz, Paul. 2004. « School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. » *Journal of Development Economics* 74 (1) : 199–250.

CHAPITRE 5

Modèle de discontinuité de la régression

Les programmes sociaux utilisent souvent un indice pour déterminer quels sont les individus ou ménages éligibles. Par exemple, les programmes de lutte contre la pauvreté ciblent généralement les ménages pauvres en les identifiant avec un indice ou un score de pauvreté. Un score de pauvreté peut se baser sur une formule de type « proxy mean » qui mesure un ensemble d'actifs du ménage. Les ménages avec de bas scores sont classés parmi les ménages pauvres et ceux dont les scores sont plus élevés sont considérés comme des ménages relativement aisés. Les responsables de programme fixent en général un seuil ou un score limite au-dessous duquel les ménages sont considérés comme pauvres et éligibles pour un programme. Le programme mexicain Progresa (Buddelmeyer et Skoufias 2004) ou le système colombien de sélection des bénéficiaires des programmes sociaux baptisé SISBEN (Barrera-Osorio, Linden et Usquiola, 2007) utilisent de telles méthodes.

Les programmes de retraite ciblent eux aussi les individus en fonction d'un indice d'éligibilité, bien qu'il soit d'un autre type. L'âge constitue un indice continu et l'âge de départ à la retraite est le seuil qui détermine l'éligibilité. Autrement dit, seules les personnes ayant dépassé un certain âge ont le droit de recevoir une retraite. Les résultats aux examens sont un autre exemple d'indice d'éligibilité continu. De nombreux pays octroient des bourses d'études ou des prix aux meilleurs élèves à un examen standardisé dont les résultats sont classés par ordre croissant. Si le nombre de bourses est limité, seuls les étudiants avec une note au-delà d'un certain seuil (par exemple la première tranche de 15 %) seront éligibles.

Concept clé :
Le modèle de discontinuité de la régression convient aux programmes qui utilisent un indice continu pour classifier les participants potentiels et un seuil pour distinguer les bénéficiaires des non-bénéficiaires.

Le modèle de discontinuité de la régression est une méthode d'évaluation d'impact qui convient aux programmes pour lesquels un indice d'éligibilité continu est établi et un seuil est clairement défini pour distinguer les bénéficiaires des non bénéficiaires. Deux conditions doivent être réunies pour pouvoir appliquer le modèle de discontinuité de la régression :

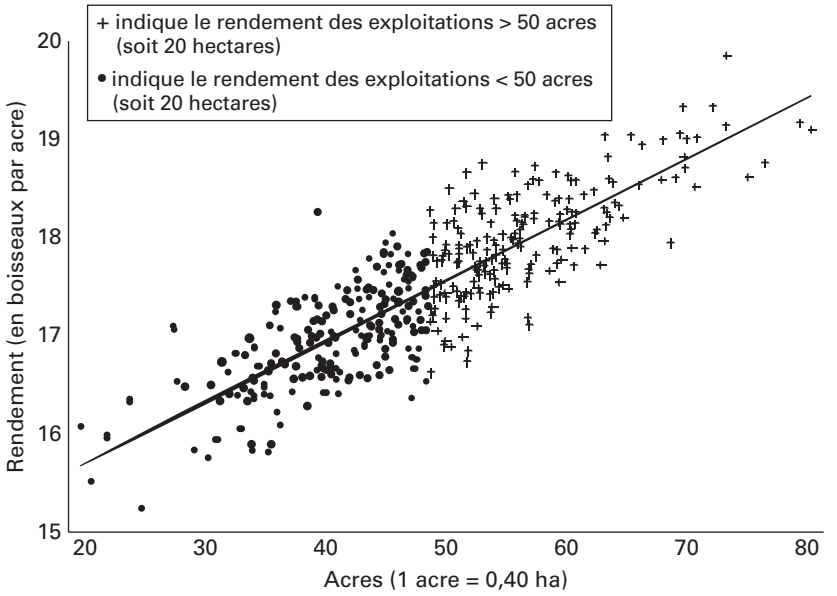
1. Un indice d'éligibilité continu doit exister, à savoir un indicateur continu permettant de classer la population à l'étude, comme un indice de pauvreté, les résultats à un examen ou l'âge.
2. Un seuil d'éligibilité doit être clairement défini, déterminant un niveau de l'indice au-dessus ou au-dessous duquel la population est considérée comme éligible au programme. Par exemple, les ménages dont l'indice de pauvreté est inférieur à 50 sur 100 peuvent être considérés comme pauvres, les personnes de 67 ans et plus peuvent être considérées comme des retraités et les étudiants obtenant un résultat de 90 sur 100 ou plus peuvent être éligibles à une bourse. Dans ces exemples, les seuils sont fixés à 50, 67 et 90 respectivement.

Cas 1 : subvention des engrais pour la riziculture

Prenons l'exemple d'un programme agricole qui subventionne les achats d'engrais par les riziculteurs dans le but d'améliorer les rendements. Le programme cible les petites et moyennes exploitations, définies au titre du programme comme des exploitations d'une superficie totale de moins de 50 acres[1]. Avant la mise en œuvre du programme, la relation entre la taille de l'exploitation et la production totale de riz est illustrée dans le graphique 5.1, les petites exploitations ayant une production totale inférieure à celle des grandes exploitations. Le seuil d'éligibilité dans ce cas est le nombre d'acres exploités, qui est fixé à 50. Conformément aux règles d'éligibilité au programme, les exploitations de moins de 50 acres sont en droit de recevoir une subvention pour l'achat d'engrais, et les exploitations de plus de 50 acres ne peuvent pas en bénéficier. Dans ce cas, il est probable que plusieurs exploitations de 48, 49 ou même 49,9 acres participent au programme. Un autre groupe d'exploitations de 50, 50,1 ou 50,2 acres sera de facto exclu du programme parce qu'elles dépassent le seuil d'éligibilité. Les exploitations de 49,9 acres ressemblent vraisemblablement en de nombreux points à celles de 50,1 acres, mais les premières reçoivent une subvention pour l'achat d'engrais au contraire des secondes. Plus nous nous éloignons du seuil d'éligibilité et plus les différences s'accentuent entre les entités éligibles et les unités non éligibles. Nous disposons toutefois d'une mesure de ces différences, les critères d'éligibilité, que nous pouvons prendre en compte.

Une fois que le programme est mis en œuvre et que les subventions sont distribuées aux petites et moyennes exploitations, les évaluateurs du programme peuvent utiliser la méthode de discontinuité de la régression pour mesurer son impact. Cette méthode mesure la différence de résultats enregistrés après l'intervention,

Figure 5.1 Rendement rizicole



comme le rendement total, pour les entités qui se situent près du seuil d'éligibilité, soit 50 acres dans notre exemple. Les exploitations légèrement trop importantes pour participer au programme constituent le groupe de comparaison et génèrent une estimation du résultat contrefactuel pour les exploitations du groupe de traitement qui sont juste au-dessous du seuil d'éligibilité. Étant donné que ces deux groupes d'exploitations étaient très similaires avant le programme et qu'ils sont exposés aux mêmes facteurs (tels que le climat, les fluctuations des cours, les politiques agricoles locales et nationales, etc.), le programme constitue la seule raison pouvant expliquer les différences de résultats après l'intervention.

La méthode de discontinuité de la régression permet d'estimer correctement l'impact d'un programme sans exclure d'unités éligibles. Il convient toutefois de noter que l'impact estimé ne s'applique qu'aux unités se situant autour du seuil d'éligibilité. Dans notre exemple, nous obtenons une estimation valide de l'impact du programme de subvention de l'achat d'engrais pour des exploitations dont la superficie est légèrement inférieure à 50 acres. L'évaluation d'impact ne permettra pas nécessairement de déterminer directement l'impact du programme sur les petites exploitations (de un ou deux acres par exemple) pour lesquelles l'impact du subventionnement des engrais pourrait être nettement différent des effets observés pour les exploitations de 48 ou 49 acres. Il n'existe pas de groupe de comparaison pour les petites exploitations étant donné qu'elles sont toutes éligibles au programme. La seule comparaison valable concerne les exploitations proches du seuil d'éligibilité de 50 acres.

Cas 2 : transferts monétaires

Supposons que nous tentions d'évaluer l'impact d'un programme de transferts monétaires sur les dépenses alimentaires journalières de ménages pauvres. Supposons également que nous puissions utiliser un indice de pauvreté[2] qui synthétise les données sur les actifs des ménages pour obtenir un score entre zéro et 100 permettant de classer les ménages des plus pauvres aux plus riches. Au départ, il est probable que, en moyenne, les ménages les plus pauvres dépensent moins en alimentation que les ménages les plus riches. La figure 5.2 représente une relation potentielle entre l'indice de pauvreté et les dépenses alimentaires journalières des ménages (le résultat).

Supposons maintenant que le programme cible uniquement les ménages pauvres définis comme ceux qui ont un indice de pauvreté inférieur à 50. Autrement dit, l'indice de pauvreté détermine l'éligibilité : le programme sera offert uniquement aux ménages qui affichent un score de 50 ou moins. Les ménages dont le score est supérieur à 50 ne sont pas éligibles. Dans cet exemple, l'indice de pauvreté constitue un indice continu avec un seuil d'éligibilité fixé à 50. La relation entre l'indice d'éligibilité et la variable de résultat (les dépenses alimentaires quotidiennes) est illus-

Figure 5.2 Dépenses des ménages et niveau de pauvreté (avant l'intervention)

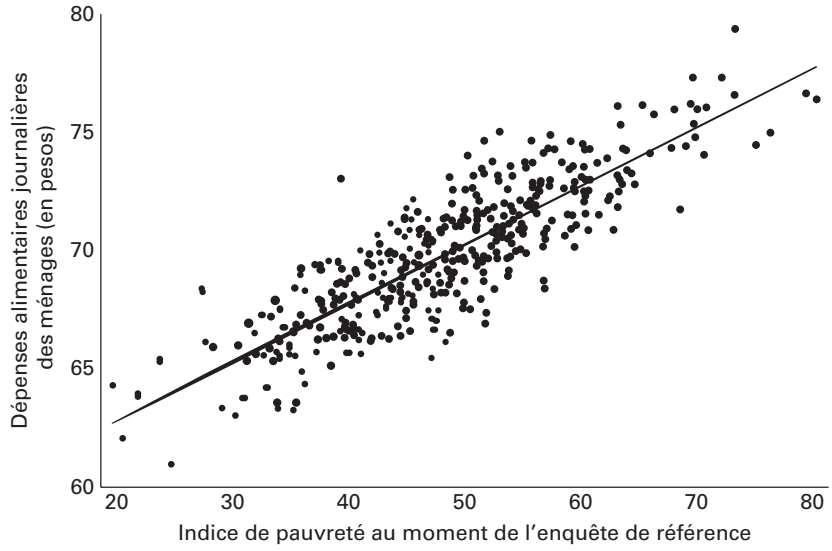
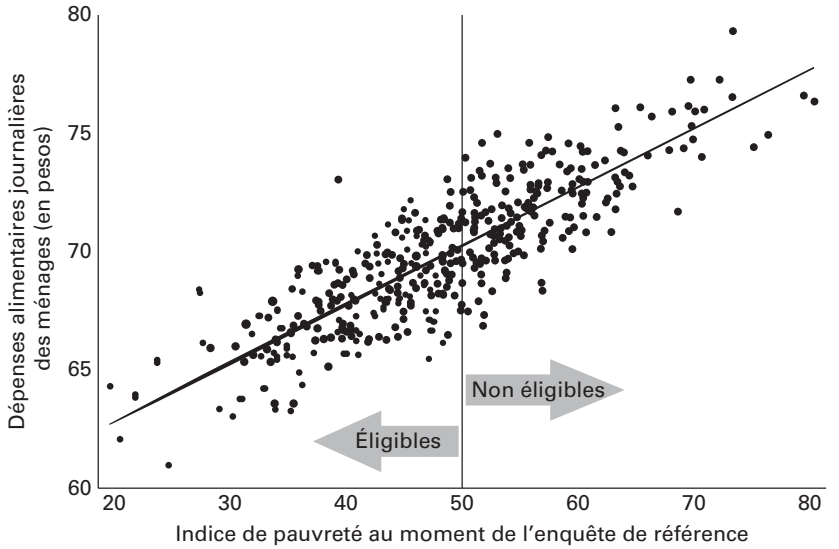


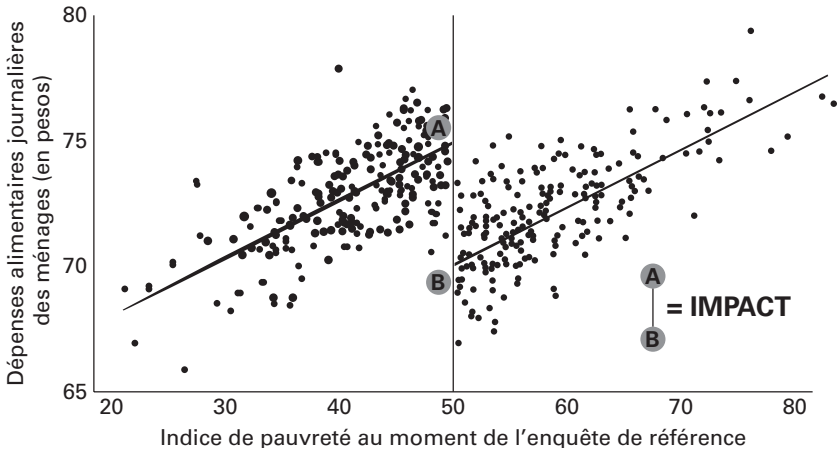
Figure 5.3 Seuil d'éligibilité au programme de transferts monétaires



trée à la figure 5.3. Les ménages se situant juste au-dessous du score limite sont éligibles au programme tandis que ceux qui se situent juste au-dessus ne le sont pas, même si ces deux types de ménages sont très similaires.

Le modèle de discontinuité de la régression utilise la discontinuité observée autour du seuil d'éligibilité pour estimer le contrefactuel. Intuitivement, nous pouvons considérer que les ménages dont le score est juste au-dessous du seuil d'éligibilité (50 et un peu moins) sont très similaires à ceux dont le score est juste au-dessus du seuil d'éligibilité (51, par exemple). Les responsables du programme ont choisi un point particulier sur l'indice continu de pauvreté (50) pour créer une coupure, ou une discontinuité, dans l'éligibilité au programme. Étant donné que les ménages qui se situent juste au-dessus du seuil des 50 sont très similaires à ceux qui sont juste en dessous, à la différence près qu'ils ne bénéficient pas des transferts monétaires, ils peuvent être utilisés comme groupe de comparaison pour les ménages qui se situent juste au-dessous du seuil d'éligibilité. Autrement dit, les ménages non éligibles au programme, mais proches du seuil d'éligibilité seront utilisés comme groupe de comparaison pour estimer le contrefactuel (à savoir les changements enregistrés dans le groupe de ménages éligibles en l'absence du programme).

Figure 5.4 Dépenses des ménages et niveau de pauvreté (après l'intervention)

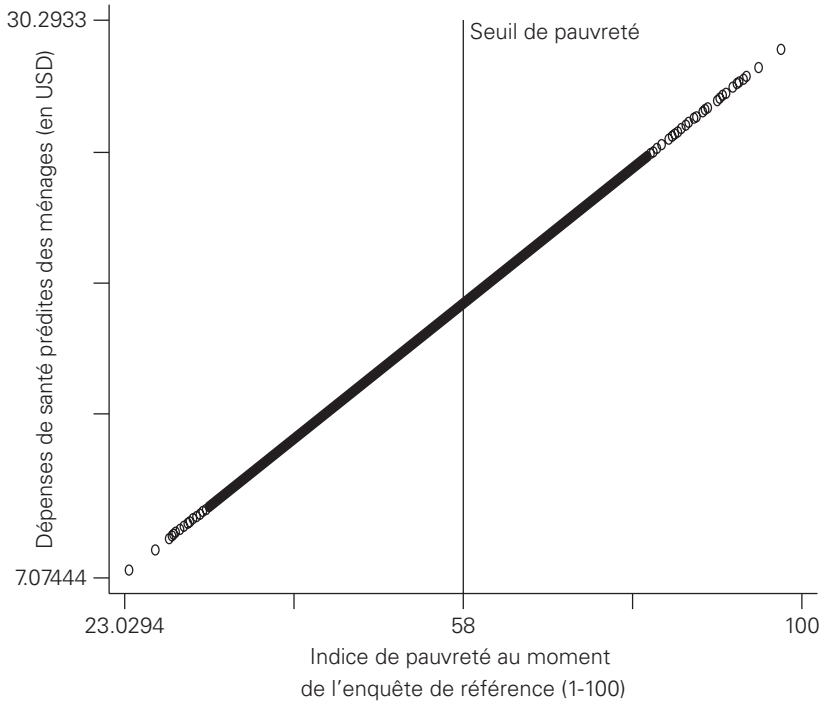


La figure 5.4 présente une situation après l'intervention qui illustre intuitivement la stratégie d'identification par discontinuité de la régression. Les résultats moyens des ménages (éligibles) dont le niveau de pauvreté au moment de l'enquête de référence est inférieur au seuil d'éligibilité sont désormais plus élevés que les résultats moyens des ménages (non éligibles) dont le niveau de pauvreté de référence était légèrement supérieur au seuil d'éligibilité. Étant donné la relation continue entre les niveaux de pauvreté et les dépenses alimentaires journalières observée avant le lancement du programme, l'existence du programme de transferts monétaires est la seule explication possible de la discontinuité constatée après l'intervention. En d'autres termes, puisque les ménages se situant dans les environs immédiats du seuil d'éligibilité (à droite et à gauche) bénéficient de caractéristiques de départ similaires, l'écart entre les dépenses alimentaires moyennes des deux groupes après l'intervention correspond à l'impact du programme.

Utilisation du modèle de discontinuité de la régression pour évaluer le Programme de subvention de l'assurance maladie (PSAM)

Appliquons maintenant le modèle de discontinuité de la régression au programme de subvention de l'assurance maladie (PSAM). Après des analyses supplémentaires sur le fonctionnement du PSAM, vous concluez que, dans la pratique, les autorités ont ciblé le programme sur les ménages au revenu inférieur au seuil national de pauvreté. Le seuil de pauvreté est fondé sur un indice qui attribue à chaque ménage du pays un score compris entre 20 et 100 en fonction de leurs

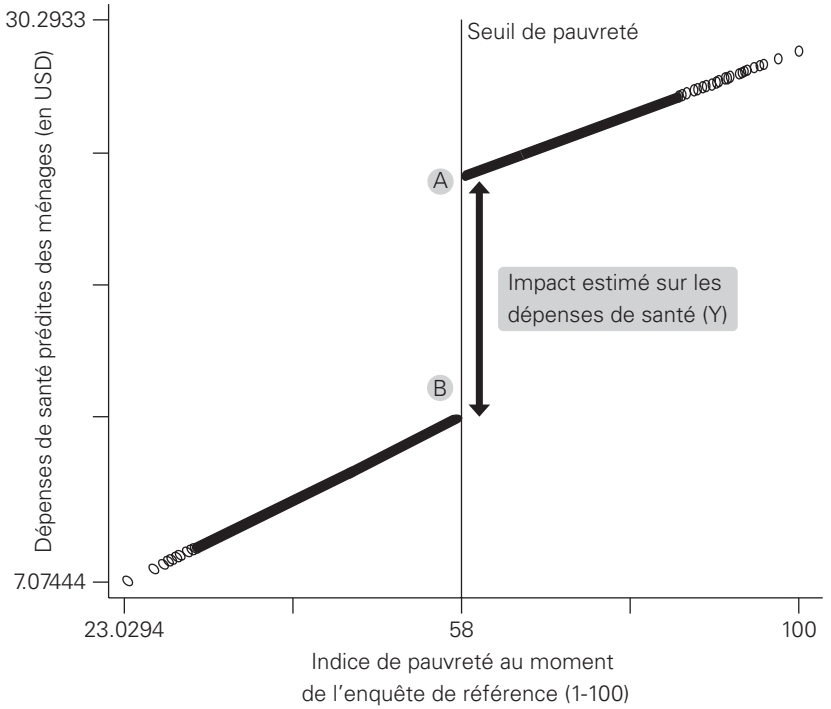
Figure 5.5 Indice de pauvreté et dépenses de santé avant le lancement du Programme de subvention de l'assurance maladie



actifs, leurs conditions de logement et leur structure sociodémographique. Le seuil de pauvreté a été officiellement fixé à 58, ce qui veut dire que tous les ménages ayant un score inférieur à 58 sont considérés comme pauvres et que tous les ménages ayant un score supérieur à 58 sont considérés comme non pauvres. Même dans les villages de traitement, seuls les ménages pauvres étaient éligibles au PSAM. Toutefois, votre échantillon comprend des données à la fois sur les ménages pauvres et sur les ménages non pauvres de ces villages.

En utilisant les ménages des villages de traitement de votre échantillon, un collègue vous aide à effectuer une régression multivariée pour établir la corrélation entre l'indice de pauvreté et les dépenses de santé prédites des ménages avant le lancement du PSAM (figure 5.5). La figure montre clairement qu'au fur et à mesure que le score de pauvreté d'un ménage augmente, la régression prédit un niveau de dépenses de santé plus élevé, ce qui indique que les ménages plus aisés ont tendance à consacrer davantage de dépenses aux médicaments et aux services de santé primaires. Il convient de noter que la relation entre l'indice de pauvreté et les dépenses de santé est continue, c'est-à-dire qu'il n'y a pas de signe de changement dans la relation autour du seuil de pauvreté.

Figure 5.6 Indice de pauvreté et dépenses de santé – deux ans après le lancement du Programme de subvention de l'assurance maladie



Deux ans après le lancement du pilote, vous constatez que seuls les ménages affichant un score inférieur à 58 (à gauche du seuil de pauvreté) ont pu participer au PSAM. À l'aide de données de suivi, vous tracez à nouveau la relation entre les scores de pauvreté et les dépenses de santé prédites (voir figure 5.6). Cette fois-ci, la relation entre l'indice de pauvreté et les dépenses de santé prédites n'est plus continue. Il y a une variation nette, ou « discontinuité » au seuil de pauvreté.

Tableau 5.1 Cas 5— Impact du PSAM selon le modèle de discontinuité de la régression (analyse de régression)

	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–9,05** (0,43)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

La discontinuité illustre une baisse des dépenses de santé de la part des ménages éligibles au programme. Étant donné que les ménages de part et d'autre du seuil de 58 sont très similaires, la seule explication possible pour la différence des dépenses de santé est l'éligibilité au programme de l'un des groupes de ménages. Vous estimez cet écart au moyen d'une régression dont les conclusions figurent dans le tableau 5.1.

QUESTION 5

A. Le résultat indiqué dans le tableau 5.1 est-il valide pour tous les ménages éligibles ?
B. Par rapport à l'impact estimé en utilisant l'assignation aléatoire, que nous indique le résultat sur les ménages dont le niveau de pauvreté est juste au-dessous de 58 ?
C. Au vu de ce résultat pour le cas 5, le PSAM doit-il être étendu à tout le pays ?

Le modèle de discontinuité de la régression en pratique

Le modèle de discontinuité de la régression a été utilisé dans différents contextes. Lemieux et Milligan (2005) analysent les effets de l'aide sociale sur l'offre de main-d'œuvre au Québec. Martinez (2004) étudie l'impact des retraites sur la consommation en Bolivie. Filmer et Schady (2009) évaluent l'impact d'un pro-

Encadré 5.1 : Aide sociale et offre de main-d'œuvre au Canada

Dans l'une des études classiques utilisant le modèle de discontinuité de la régression, les auteurs examinent une discontinuité nette dans un programme d'assistance sociale au Québec (Canada), pour comprendre l'impact du programme sur des indicateurs d'insertion professionnelle. Ce programme d'assistance, financé par le Régime d'assistance publique du Canada, vient en aide aux chômeurs. Pendant de nombreuses années, le programme a versé des montants nettement inférieurs aux individus de moins de 30 ans sans enfants en comparaison aux personnes de plus de 30 ans (185 dollars par mois contre 507 dollars).

Afin d'évaluer rigoureusement ce programme, Lemieux et Milligan (2005) limitent leur échantillon aux hommes sans enfants et sans diplômes d'éducation secondaire et utilisent des données du recensement canadien et de l'Enquête sur la population active. Pour justifier le choix de l'approche de discontinuité de la régression, ils démontrent que les hommes proches du seuil de discontinuité (entre 25 et 39 ans) présentent des caractéristiques observables très similaires.

En comparant les sujets des deux côtés du seuil d'éligibilité, les auteurs montrent que l'accès à des prestations sociales plus élevées réduit d'environ 4,5 % le taux d'insertion professionnelle des hommes sans enfants de cette tranche d'âges.

Source : Lemieux et Milligan, 2005.

gramme d'octroi de bourses aux étudiants pauvres sur la scolarisation et les résultats scolaires au Cambodge. Buddelmeyer et Skoufias (2004) comparent la performance de la discontinuité de la régression à celle de l'assignation aléatoire dans le cas du programme Progresa et concluent que les impacts estimés à l'aide de ces deux méthodes sont similaires pour une grande majorité des résultats analysés. Certains de ces exemples sont décrits plus en détail dans les encadrés 5.1, 5.2 et 5.3.

Encadré 5.2 : Frais de scolarité et taux de scolarisation en Colombie

En Colombie, Barrera-Osorio, Linden et Urquiola (2007) utilisent le modèle de discontinuité de la régression pour évaluer l'impact d'un programme de réduction des frais de scolarité (Gratuidad) sur les taux de scolarisation à Bogota. La population cible du programme est définie à l'aide de l'indice SISBEN, un indice de pauvreté continu dont la valeur est déterminée en fonction de caractéristiques des ménages comme l'emplacement et les matériaux de construction du logement, les services disponibles, les données démographiques, l'état de santé, le niveau d'éducation, le niveau de revenus et le travail exercé par les membres du ménage. Le gouvernement définit deux seuils sur l'indice SISBEN : les enfants des ménages dont le score est inférieur au seuil n° 1 sont éligibles pour le programme d'éducation gratuite de la 1ère à la 11ème année ; les enfants des ménages dont le score est compris entre le seuil n° 1 et le seuil n° 2 sont éligibles à une subvention de 50 % des frais de scolarité pour la 10ème et 11ème année ; et les enfants des ménages dont le score est supérieur au seuil n° 2 ne sont pas éligibles pour le programme d'éducation gratuite ou de subventions.

Les auteurs utilisent le modèle de discontinuité de la régression pour quatre raisons. Premièrement, les caractéristiques des ménages comme le niveau de revenus ou d'éducation du chef de famille sont continues tout au long de l'indice SISBEN au moment de l'enquête de référence, il n'y a donc pas de « bond » dans les caractéristiques le long de l'indice. Deuxièmement, les ménages de part et d'autre des seuils définis présentent des caractéristiques similaires, ce qui indique que l'approche a créé des groupes de comparaison crédibles. Troisièmement, un grand échantillon de ménages est disponible. Enfin, le gouvernement n'a pas révélé la formule utilisée pour calculer l'indice SISBEN pour que les ménages ne puissent pas manipuler leurs scores.

En utilisant le modèle de discontinuité de la régression, les chercheurs découvrent que le programme a un impact positif significatif sur les taux de scolarisation. Ainsi, le taux de scolarisation augmente de trois points pour les élèves d'écoles primaires provenant de ménages situés au-dessous du seuil n° 1 et de six points pour les lycéens venant de ménages se situant entre les seuils n° 1 et n° 2. Cette étude démontre les avantages de la réduction des frais de scolarité directs, en particulier pour les étudiants à risque. Toutefois, les auteurs appellent également à poursuivre les recherches sur l'élasticité-prix afin de perfectionner l'élaboration des programmes de subventions comme celui-ci.

Source: Barrera-Osorio, Linden et Urquiola 2007.

Encadré 5.3 : Filets de protection sociale fondés sur un indice de pauvreté en Jamaïque

Le modèle de discontinuité de la régression est également utilisé pour évaluer l'impact d'un filet de protection sociale en Jamaïque. En 2001, le Gouvernement jamaïcain lance le programme PATH (Programme of Advancement through Health and Education) afin de renforcer les investissements dans le capital humain et d'améliorer le ciblage de l'aide sociale aux pauvres. Le programme offre des allocations de santé et d'éducation aux enfants provenant de ménages pauvres éligibles à condition qu'ils soient scolarisés et qu'ils effectuent des visites médicales régulières. L'allocation mensuelle moyenne par enfant s'élève à environ 6,50 $, auxquels vient s'ajouter l'exonération de certains frais de soins de santé et de scolarité.

Étant donné que l'éligibilité au programme est déterminée par un score, Levy et Ohls (2007) comparent les ménages se situant juste au-dessous du seuil d'éligibilité et ceux juste au-dessus (entre 2 et 15 points du seuil). Les chercheurs justifient l'utilisation du modèle de discontinuité de la régression à partir de données de référence indiquant que les ménages du groupe de traitement et ceux du groupe de comparaison présentent un niveau de pauvreté similaire, sur la base d'un score « proxy mean », et un degré de motivation similaire, tous les ménages de l'échantillon ayant demandé à bénéficier du programme. Les chercheurs utilisent aussi le score d'éligibilité au programme dans l'analyse de régression multivariée pour contrôler pour d'éventuelles différences observées entre les deux groupes.

Levy et Ohls (2007) découvrent que le programme PATH entraîne une hausse de la scolarisation des enfants de six à 17 ans de 0,5 jour par mois en moyenne, un résultat satisfaisant étant donné que le taux de scolarisation de départ était relativement élevé, à 85 %. Par ailleurs, ils constatent une augmentation d'environ 38 % du nombre de visites médicales pour les enfants de zéro à six ans. Bien que les chercheurs ne puissent pas déceler d'impacts à long terme sur les résultats scolaires ou les indicateurs de santé, ils concluent que la magnitude des impacts identifiés concorde, globalement, avec les programmes de transferts monétaires conditionnels mis en œuvre dans d'autres pays. Enfin, cette évaluation est fondée sur des données quantitatives et qualitatives collectées par des systèmes d'information, des entretiens, des groupes focaux et des enquêtes auprès des ménages.

Source: Levy and Ohls 2007.

Limites et interprétation du modèle de discontinuité de la régression

Le modèle de discontinuité de la régression estime l'impact moyen *local* aux alentours du seuil d'éligibilité, c'est-à-dire au point où le groupe de traitement et le groupe de comparaison sont les plus similaires. En s'approchant du seuil, les unités qui se situent à gauche et à droite du seuil sont de plus en plus similaires. En fait, dans la proximité immédiate du seuil d'éligibilité, les unités de part et d'autre du seuil sont tellement similaires que la comparaison est aussi précise qu'en utilisant l'assignation aléatoire pour générer les groupes de traitement et un de comparaison.

Le modèle de discontinuité de la régression évalue l'impact du programme *localement* aux alentours du seuil d'éligibilité. L'estimation ne peut pas systématiquement être généralisée aux unités dont le score est plus éloigné de ce seuil, c'est-à-dire aux parties de la distribution où les unités éligibles et non éligibles ne sont plus similaires. Le fait que cette approche ne permette pas de calculer l'effet moyen du traitement pour tous les participants au programme peut être considéré comme un avantage ou un inconvénient en fonction de l'information recherchée. Si l'évaluation vise principalement à savoir si le programme devrait ou non être mis en œuvre, l'effet moyen du traitement sur l'ensemble de la population éligible est probablement le paramètre le plus pertinent, et l'impact local estimé par le modèle de discontinuité de la régression n'est pas satisfaisant. Toutefois, si la question est de savoir si le programme doit être réduit ou au contraire élargi, le modèle de discontinuité de la régression fournit précisément l'impact local utile pour prendre cette décision.

Le fait que cette méthode évalue les effets locaux moyens du traitement représente également un défi en termes de puissance statistique de l'analyse. Étant donné que les effets ne sont mesurés qu'autour du seuil d'éligibilité, cette méthode utilise moins d'observations que d'autres méthodes utilisant toutes les unités disponibles. Le modèle de discontinuité de la régression requiert des échantillons d'évaluation relativement importants pour obtenir une puissance statistique suffisante. Dans la pratique, il faut déterminer une bande autour du seuil d'éligibilité sur laquelle portera l'évaluation en assurant l'équilibre des caractéristiques observées des populations au-dessus et au-dessous du seuil d'éligibilité. Il est ensuite possible de répéter l'estimation avec des bandes différentes pour vérifier si les résultats sont robustes au changement de la bande considérée. En règle générale, plus la bande est large, plus la puissance statistique est élevée puisqu'un plus grand nombre d'observations sont prises en compte. Toutefois, en s'éloignant du seuil d'éligibilité, il peut être nécessaire de formuler certaines hypothèses concernant les formes fonctionnelles pour obtenir une estimation crédible de l'impact.

L'autre réserve concernant le modèle de discontinuité de la régression vient du fait que la spécification peut varier en fonction de la forme fonctionnelle utilisée pour modéliser la relation entre l'indice d'éligibilité et le résultat. Dans l'exemple du programme de transferts monétaires, nous avons supposé que la relation entre l'indice de pauvreté des ménages et leurs dépenses alimentaires journalières était simple et linéaire au moment de l'enquête de référence. En réalité, la relation entre l'indice d'éligibilité et le résultat (Y) au moment de l'enquête de référence peut être beaucoup plus complexe et comprendre des relations et des interactions non linéaires. Si l'estimation ne tient pas compte de ces relations complexes, elles risquent d'être interprétées comme un signe de discontinuité dans les résultats recueillis après l'intervention. Dans la pratique, l'impact du programme peut être estimé en utilisant plusieurs formes fonctionnelles (linéaire, quadratique, cubique, etc.) pour déterminer si les estimations de l'impact sont robustes aux changements de la forme fonctionnelle.

Même en tenant compte de ces réserves, le modèle de discontinuité de la régression permet d'obtenir des estimations non biaisées de l'impact du programme aux alentours du seuil d'éligibilité. Cette approche se base sur des indices d'éligibilité

continus et des règles d'allocation de programme qui sont fréquemment utilisés dans les programmes sociaux. Lorsqu'un ciblage basé sur un indice est utilisé, il n'est pas nécessaire d'exclure du programme un groupe de ménages ou de personnes éligibles pour réaliser l'évaluation puisque le modèle de discontinuité de la régression peut être utilisé à la place.

Notes

1. 1 acre = 0,40 ha.
2. Ceci est souvent appelé un test « proxy mean » parce qu'il utilise les actifs du ménage comme indicateurs pour approximer les moyens ou le pouvoir d'achat du ménage.

Références

Barrera-Osorio, Felipe, Leigh Linden et Miguel Urquiola. 2007. « The Effects of User Fee Reductions on Enrollment: Evidence from a Randomized Natural Experiment. » Columbia University et Banque mondiale, Washington, DC.

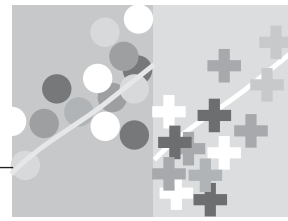
Buddelmeyer, Hielke et Emmanuel Skoufias. 2004. « An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA. » Document de travail consacré à la recherche sur les politiques 3386, IZA Discussion Paper 827, Banque mondiale, Washington, DC.

Filmer, Deon et Norbert Schady. 2009. « School Enrollment, Selection and Test Scores. » Document de travail consacré à la recherche sur les politiques 4998, Banque mondiale, Washington, DC.

Lemieux, Thomas et Kevin Milligan. 2005. « Incentive Effects of Social Assistance: A Regression Discontinuity Approach. » NBER Working Paper 10541, National Bureau of Economic Research, Cambridge, MA.

Levy, Dan et Jim Ohls. 2007. « Evaluation of Jamaica's PATH Program: Final Report. » Mathematica Policy Research, Inc., Ref. 8966-090, Washington, DC.

Martinez, S. 2004. « Pensions, Poverty and Household Investments in Bolivia. » University of California, Berkeley, CA.



CHAPITRE 6

Double différence

Les trois méthodes d'évaluation d'impact abordées jusqu'à présent (l'*assignation aléatoire*, la *promotion aléatoire du traitement* et le *modèle de discontinuité de la régression*) permettent d'estimer le contrefactuel sur la base de règles d'allocation des programmes qui sont connues et comprises par l'évaluateur. Nous avons exposé les raisons pour lesquelles ces méthodes fournissent des estimations crédibles du contrefactuel en utilisant relativement peu d'hypothèses et conditions. Les deux méthodes que nous allons maintenant aborder (la *double différence* et l'*appariement*) pourvoient l'évaluateur d'outils utilisables lorsque les règles d'assignation des programmes sont moins claires ou lorsqu'aucune des autres méthodes décrites ci-dessus n'est applicable. Comme nous allons le voir, la double différence (DD) et l'appariement constituent de puissants outils statistiques qui sont souvent utilisés ensemble ou en conjonction avec d'autres méthodes d'évaluation d'impact.

Tant la double différence que l'appariement sont couramment utilisés, mais reposent sur des hypothèses plus contraignantes que les méthodes de sélection aléatoire. Précisons tout de suite que ces deux méthodes ne peuvent pas être appliquées sans des données de référence collectées avant le début du programme à évaluer[1].

Comme son nom l'indique, la méthode de la double différence compare les *différences* de résultats au fil du temps entre une population participant à un programme (le groupe de traitement) et une autre n'y participant pas (le groupe de comparaison). Prenons l'exemple d'un programme de construction de routes qui ne peut pas faire l'objet d'une assignation aléatoire ni d'une attribution sur la base d'un indice continu assorti d'un seuil d'éligibilité, rendant l'utilisation du modèle de discontinuité de la régression impossible. Comme l'un des objectifs de ce programme est d'améliorer l'accès au marché du travail, le taux d'emploi

Concept clé :
La méthode de la double différence estime le contrefactuel pour le changement du résultat dans le groupe de traitement en utilisant le changement du résultat dans le groupe de comparaison. Cette méthode permet de prendre en compte les différences entre le groupe de traitement et le groupe de comparaison qui sont invariables dans le temps.

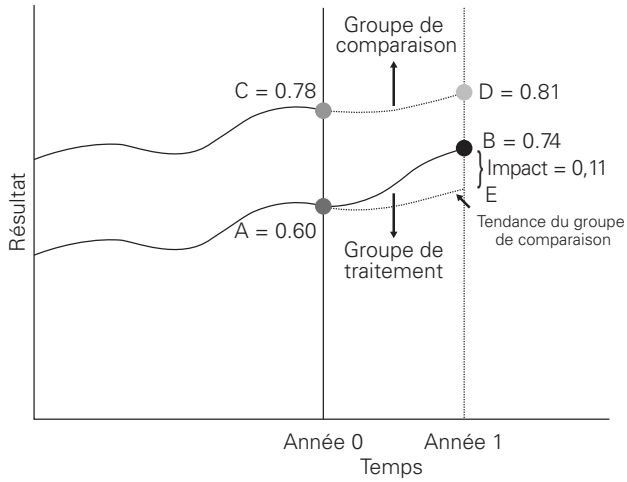
constitue l'un des indicateurs de résultat. Comme nous l'avons vu au chapitre 3, la simple observation du changement du taux de chômage avant et après la mise en œuvre du programme ne suffit pas à mesurer son effet causal. En effet, de nombreux autres facteurs variables dans le temps peuvent influencer le taux de chômage. De même, comparer les régions qui ont reçu le programme à celles qui ne l'ont pas reçu serait problématique puisqu'il peut exister des raisons non observées pour lesquelles certaines régions ont bénéficié du programme et d'autres non (il s'agit du problème du biais de sélection évoqué dans la comparaison avec-sans, ou inscrits et non inscrits).

Cependant, que se passerait-il si nous combinions les deux méthodes pour comparer les résultats avant-après d'un groupe qui a pris part au programme et d'un groupe qui n'y a pas pris part ? La différence dans les résultats avant-après pour le groupe participant (la première différence) *contrôle* pour les facteurs invariables dans le temps qui affectent ce groupe, pour la simple raison que nous comparons le groupe à lui-même. La différence avant-après ne tient toutefois pas compte des facteurs externes variables dans le temps. Une manière de prendre en compte ces facteurs externes variables dans le temps est de mesurer la différence de résultats avant-après pour un groupe qui *n'a pas* participé au programme, mais qui a été exposé aux mêmes conditions externes (la deuxième différence). Si nous épurons la première différence des effets des autres facteurs variables dans le temps qui influent sur les résultats en soustrayant la deuxième différence, nous éliminons la principale source de biais qui posait problème dans la simple comparaison avant-après. La double différence combine donc les deux contrefactuels contrefaits (comparaisons avant-après et comparaisons avec-sans entre les participants et les non participants) pour produire une meilleure estimation du contrefactuel. Dans le cas de notre programme routier, la méthode DD comparerait par exemple la différence entre les taux d'emploi observés dans les zones concernées par la construction des routes avant et après la mise en œuvre du programme, et ce même changement dans le taux d'emploi observé dans les zones où le programme n'a pas été mis en œuvre.

Il est important de relever que le contrefactuel estimé par la méthode de double différence correspond au *changement* des résultats pour le groupe de comparaison. Le groupe de traitement et le groupe de comparaison ne doivent pas nécessairement être similaires avant l'intervention. Toutefois, pour que la méthode DD soit valide, le groupe de comparaison doit fournir une estimation précise du changement de résultats qui aurait prévalu dans le groupe de traitement s'il n'avait pas participé au programme. Pour appliquer la double différence, il suffit de mesurer les résultats du groupe de participants (le groupe de traitement) et ceux du groupe de non participants (le groupe de comparaison) tant avant qu'après la mise en œuvre du programme. La méthode ne requiert pas de préciser les règles d'assignation du programme.

La figure 6.1 illustre la méthode de la double différence. Un groupe de traitement participe à un programme et un groupe de comparaison n'y participe pas.

Figure 6.1 Double différence



Les variables de résultats avant et après pour le groupe de traitement sont *A* et *B* respectivement tandis que le résultat du groupe de comparaison passe de *C* avant le programme à *D* après sa mise en œuvre.

Souvenez-vous des deux contrefactuels contrefaits : la différence de résultats avant et après l'intervention pour le groupe de traitement (*B* – *A*) et la différence de résultats[2] après l'intervention entre le groupe de traitement et le groupe de comparaison (*B* – *D*). Selon la méthode de la double différence, l'estimation du contrefactuel est obtenue en calculant la différence des résultats du groupe de comparaison avant et après l'intervention (*D* – *C*). Ce contrefactuel pour le changement du résultat à travers le temps est ensuite soustrait du changement du résultat observé pour le groupe de traitement (*B* – *A*).

En résumé, l'impact du programme est calculé comme la différence entre deux différences :

Impact par DD = (*B* – *A*) – (*D* – *C*) = (*B* – *E*) = (0,74 – 0,60) – (0,81 – 0,78) = 0,11.

Le contenu de la figure 6.1 peut également être illustré sous forme de tableau. Le tableau 6.1 expose les composantes de l'estimation par double différence. La première ligne du tableau contient les résultats du groupe de traitement, avant (*A*) et après (*B*) l'intervention. La comparaison avant-après pour le groupe de traitement constitue la première différence (*B* – *A*). La deuxième ligne du tableau contient les résultats du groupe de comparaison avant (*C*) et après (*D*) l'intervention. La deuxième différence correspond donc à *D* – *C*.

Tableau 6.1 Double différence

	Après	Avant	Différence
Traitement/participants	*B*	*A*	*B – A*
Comparaison/ non participants	*D*	*C*	*D – C*
Différence	*B – D*	*A – C*	*DD = (B – A) – (D – C)*

	Après	Avant	Différence
Traitement/participants	0,74	0,60	*0,14*
Comparaison/ non participants	0,81	0,78	*0,03*
Différence	*–0,07*	*–0,18*	*DD = 0,14 – 0,03 = 0,11*

La méthode de la double différence calcule l'impact estimé selon la formule suivante :

1. Nous calculons la différence de résultat (Y) entre la situation avant et après pour le groupe de traitement ($B - A$).
2. Nous calculons la différence de résultat (Y) entre la situation avant et après pour le groupe de comparaison ($D \neg C$).
3. Nous calculons ensuite la différence entre la différence de résultats pour le groupe de traitement ($B - A$) et la différence pour le groupe de comparaison ($D - C$), soit DD = ($B - A$) – ($D - C$). La double différence est notre estimation d'impact.

En quoi la méthode de la double différence est-elle utile ?

Pour comprendre comment cette méthode peut être utile, reprenons le deuxième contrefactuel contrefait, qui compare les participants au programme aux non participants. Souvenez-vous que le principal problème dans ce cas est que les deux groupes ont potentiellement des caractéristiques différentes qui peuvent être à l'origine des différences de résultats entre les deux groupes. Les différences *non observées* entre les caractéristiques sont particulièrement préoccupantes : par définition, il est impossible de prendre en compte les différences de caractéristiques non observées dans l'analyse.

La méthode de la double différence contribue à résoudre ce problème dans la mesure où de nombreuses caractéristiques unitaires ou individuelles peuvent raisonnablement être considérées comme *invariables dans le temps*. Prenons l'exemple des caractéristiques *observées*, telles que l'année de naissance d'une personne, la proximité d'une région à la mer, le niveau de développement économique d'une ville ou le niveau d'éducation d'un père de famille. Même si la plupart de ces variables peuvent plausiblement influencer des résultats, elles sont peu susceptibles de changer pendant une évaluation. En suivant le même raisonnement, de nombreuses caractéristiques *non observées* peuvent elles aussi être considérées comme invariables dans le temps. Prenons par exemple l'intelligence d'une personne ou des traits de caractère comme la motivation, l'optimisme, l'autodiscipline ou les antécédents médicaux d'une famille. Il est probable qu'un grand nombre de ces caractéristiques intrinsèques n'évoluent pas avec le temps.

Lorsqu'un même individu est observé avant et après la mise en œuvre d'un programme et que nous calculons une simple différence de résultat pour ce dernier, nous annulons l'effet de toutes les caractéristiques qui sont uniques à cet individu ou qui ne changent pas avec le temps. En effet, des facteurs constants à travers le temps ne peuvent pas expliquer le changement du résultat à travers le temps. Il est important de souligner que nous contrôlons ainsi non seulement pour l'effet des caractéristiques invariables *observées* (que nous pouvons tenir en compte ou contrôler), mais aussi celui des caractéristiques invariables *non observées* comme celles mentionnées ci-dessus.

L'hypothèse des « tendances égales » dans la méthode de la double différence

La méthode de la double différence permet de tenir compte des différences entre le groupe de traitement et le groupe de comparaison qui sont invariables dans le temps ; toutefois, elle ne permet pas d'éliminer les différences entre ces deux groupes qui changent au cours du temps. Dans l'exemple du projet routier mentionné ci-dessus, si les zones d'intervention bénéficient également de la construction d'un nouveau port maritime, nous ne pourrons pas séparer l'effet de la construction de routes de l'effet de la construction du port en utilisant l'approche de la double différence. Pour que celle-ci puisse fournir une estimation valable du contrefactuel, il faut partir de l'hypothèse qu'il n'existe aucune différence variable dans le temps entre le groupe de traitement et le groupe de comparaison.

En d'autres termes, il faut partir du principe que, en l'absence du programme, les changements du résultat entre le groupe de traitement et le groupe de comparaison évolueraient en parallèle. Autrement dit, les résultats varieraient au même rythme pour les deux groupes sans le traitement, que ce soit à la hausse ou à la baisse. Il faut donc que les résultats affichent des *tendances équivalentes en l'absence de traitement*.

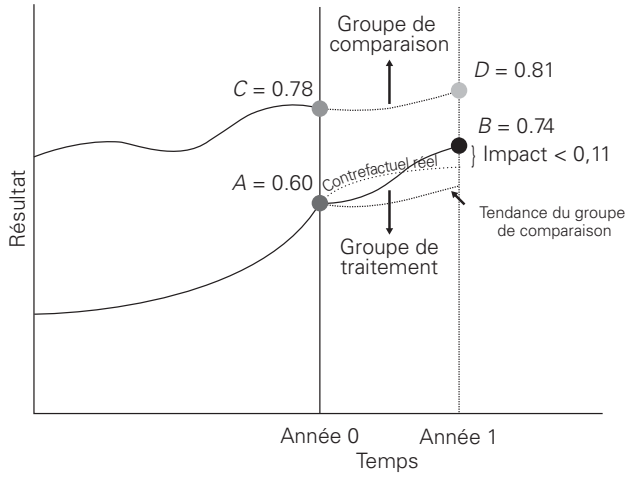
Malheureusement, nous n'avons aucun moyen de prouver que les différences entre le groupe de traitement et le groupe de comparaison auraient évolué en parallèle en l'absence du programme. En effet, nous ne pouvons pas observer la façon dont le groupe de traitement évoluerait en l'absence du traitement (à savoir le contrefactuel).

Dès lors, lorsque nous utilisons la méthode de la double différence, nous devons *postuler* qu'en l'absence du programme, le résultat du groupe de traitement aurait évolué en parallèle à celui du groupe de comparaison. La figure 6.2 illustre une violation de ce postulat fondamental qui est requis pour que la méthode de la double différence produise des estimations d'impact crédibles. Si les tendances du résultat diffèrent entre le groupe de traitement et le groupe de comparaison, l'impact estimé du traitement obtenu grâce à cette méthode sera invalide ou biaisé. En effet, dans ce cas la tendance pour le groupe de comparaison n'est pas une estimation valide de la tendance contrefactuelle qu'aurait suivie le groupe de traitement en l'absence de programme. Dans la figure 6.2, le résultat du groupe de comparaison progresse plus lentement que le résultat du groupe de traitement en l'absence du programme, donc, l'utilisation de la tendance du groupe de comparaison comme contrefactuel de la tendance du groupe de traitement entraîne une surestimation de l'impact du programme.

Tester la validité de l'hypothèse des « tendances équivalentes » dans la méthode de la double différence

La validité de l'hypothèse des tendances équivalentes peut être testée même si elle ne peut pas être totalement avérée. Une bonne approche pour tester sa validité consiste à comparer les tendances du résultat du groupe de traitement et du groupe de comparaison *avant* la mise en œuvre du programme. Si les résultats évoluent en parallèle avant le début du programme, il est probable qu'ils auraient continué à évoluer en parallèle durant la période consécutive à l'intervention.

Figure 6.2 Double différence en cas de divergence des tendances du résultat



Pour vérifier l'équivalence des tendances avant l'intervention, il faut avoir à disposition au moins deux rondes de données tant pour le groupe de traitement que le groupe de comparaison avant que le programme ne soit lancé. L'évaluation nécessite donc trois rondes de donnée : deux observations avant l'intervention pour évaluer les tendances avant le lancement du programme et au moins une observation après l'intervention pour évaluer l'impact par double différence.

Une deuxième manière de tester l'hypothèse des tendances équivalentes consiste à effectuer un test dit « placebo ». Ce test formule une estimation par double différence supplémentaire en utilisant un « faux » groupe de traitement, c'est-à-dire un groupe qui n'a en réalité pas été affecté par le programme. Par exemple, pour estimer l'impact d'un programme de tutorat personnalisé sur la probabilité que les étudiants de 7ème année fréquentent davantage l'école, vous choisissez des étudiants de 8ème année comme groupe de comparaison. Pour savoir si les élèves de 7ème et 8ème année présentent les mêmes tendances en matière de taux de fréquentation scolaire, vous pourriez analyser si les élèves de 6ème et 8ème année présentent les mêmes tendances. Vous savez que les élèves de 6ème année ne sont pas affectés par le programme ; donc si vous effectuez une estimation par double différence en utilisant des étudiants de 8ème année comme groupe de comparaison et des étudiants de 6ème année comme faux groupe de traitement, vous *devriez* obtenir un impact nul. Si ce n'est pas le cas, l'impact estimé doit provenir d'une différence sous-jacente entre les tendances de ces deux groupes d'élèves. Cela remettrait également en question l'existence de tendances équivalentes pour les étudiants de 7ème et 8ème année en l'absence de programme.

Un test placebo peut être réalisé non seulement avec un faux groupe de traitement, mais également avec un faux résultat. Dans l'exemple du tutorat, vous pouvez aussi vérifier la validité de votre choix des étudiants de 8ème année comme groupe de comparaison en évaluant l'impact du tutorat sur un résultat qui ne sera pas affecté, par exemple le nombre de frères et sœurs des étudiants. Si votre estimation par double différence conclut que le tutorat a un impact sur le nombre de frères et sœurs des étudiants, il est probable que le groupe de comparaison ne soit pas adéquat.

Il existe une quatrième manière de tester l'hypothèse des tendances équivalentes, et ce, en appliquant l'estimation par double différence à différents groupes de comparaison. Dans l'exemple du tutorat, vous pouvez effectuer dans un premier temps l'estimation en utilisant les étudiants de 8ème année comme groupe de comparaison, puis vous pouvez formuler une deuxième estimation en utilisant les étudiants de 6ème année. Si les impacts estimés dans les deux cas sont équivalents, il est probable que les deux groupes de comparaison soient valides.

Utilisation de la double différence pour évaluer le Programme de subvention de l'assurance maladie (PSAM)

La méthode de la double différence peut être utilisée pour évaluer l'impact du programme de subvention de l'assurance maladie (PSAM). Dans ce scénario, vous disposez de deux rondes de données sur deux groupes de ménages, l'un ayant participé au programme et l'autre non. Vous savez qu'en raison du biais de sélection vous ne pouvez pas effectuer une simple comparaison des dépenses de santé entre les participants et les non participants. Étant donné que vous disposez de données couvrant deux périodes pour chaque ménage de l'échantillon, vous pouvez utiliser les données pour comparer le changement des dépenses des deux groupes, en partant du principe que le changement des dépenses de santé des non participants reflète ce qu'auraient été les dépenses des participants en l'absence du programme (voir tableau 6.2). Au passage, la façon dont vous calculez la double différence dans le tableau, à savoir par colonne ou par ligne, fournit le même résultat.

Vous estimez ensuite l'impact en utilisant une analyse de régression (tableau 6.3). À l'aide d'une régression linéaire simple, vous découvrez que le programme a entraîné une réduction des dépenses de santé des ménages de 7,8 dollars. Vous affinez ensuite votre analyse en effectuant une régression linéaire multivariée pour contrôler pour plusieurs autres facteurs, et vous constatez la même réduction des dépenses de santé des ménages.

QUESTION 6

A. Quelles sont les hypothèses fondamentales qui sous-tendent le résultat du cas 6 ?
B. Au vu de ces résultats pour le cas 6, le PSAM doit-il être élargi à l'échelle nationale ?

Tableau 6.2 Cas 6— Impact du PSAM selon la méthode de la double différence (comparaison des moyennes)

	Après (suivi)	Avant (données de référence)	Différence
Inscrits	7,8	14,4	–6,6
Non-inscrits	21,8	20,6	1,2
Différence			DD = –6,6 – 1,2 = –7,8

Tableau 6.3 Cas 6— Impact du PSAM selon la méthode de la double différence (analyse de régression)

	Régression linéaire	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–7,8**	–7,8**
	(0,33)	(0,33)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

La méthode de la double différence en pratique

Malgré les limites qu'elle présente, la méthode de la double différence reste l'une des plus utilisées pour l'évaluation d'impact. Il en existe de nombreux exemples dans la littérature. Par exemple, Duflo (2001) analyse l'impact de la construction d'écoles sur la scolarisation, les indicateurs d'emploi et les salaires en Indonésie. DiTella et Schargrodsky (2005) cherchent quant à eux à savoir si un renforcement des forces de police réduit la criminalité. Un autre exemple important est exposé à l'encadré 6.1.

Encadré 6.1 : Privatisation de l'approvisionnement en eau et mortalité infantile en Argentine

Galiani, Gertler et Schargrodsky (2005) utilisent la méthode de la double différence pour déterminer si la privatisation des services d'approvisionnement en eau améliore les résultats dans le domaine de la santé et contribue à réduire la pauvreté. Dans les années 90, l'Argentine a lancé l'une des plus grandes campagnes de privatisation de son histoire, transférant le contrôle de compagnies locales d'approvisionnement en eau à des sociétés privées desservant environ 30 % des municipalités du pays et 60 % de la population. Le processus de privatisation a pris une décennie, la plus grande vague des privatisations ayant eu lieu après 1995.

Galiani, Gertler et Schargrodsky (2005) utilisent la privatisation graduelle des compagnies d'approvisionnement en eau pendant dix ans pour déterminer l'impact de cette privatisation sur la mortalité des enfants de moins de cinq ans. Avant 1995, les taux de mortalité infantile diminuent à un rythme globalement similaire dans toute l'Argentine, mais après 1995, ils baissent plus rapidement dans les municipalités où les services d'approvisionnement en eau ont été privatisés. Selon les chercheurs, l'hypothèse fondamentale sous-tendant la méthode de la double différence est probablement correcte dans ce contexte. Premièrement, la décision de privatiser les infrastructures n'est pas corrélée à des chocs économiques ou aux niveaux historiques de la mortalité infantile. Deuxièmement, les municipalités constituant le groupe de traitement et les municipalités constituant le groupe de comparaison affichent des tendances de mortalité infantile comparables avant le lancement de la privatisation.

Les chercheurs vérifient la validité de leurs conclusions en décomposant l'impact de la privatisation sur la mortalité infantile par cause de décès. Ils découvrent que la privatisation des services d'approvisionnement en eau est corrélée avec la réduction du nombre de décès liés à des maladies infectieuses et parasitaires, mais pas aux décès non liés à la qualité de l'eau (comme les accidents ou les maladies congénitales). L'évaluation permet de déterminer que la mortalité infantile baisse de près de 8 % dans les zones où les services d'approvisionnement en eau ont été privatisés, et que l'impact est plus marqué (environ 26 %) dans les zones les plus pauvres, où l'expansion du réseau de distribution d'eau a été la plus importante. Cette étude informe plusieurs débats importants sur la privatisation des services publics. Les chercheurs concluent que le secteur privé réglementé en Argentine est plus efficace que le secteur public pour améliorer les indicateurs d'accès, de services et, surtout, de mortalité infantile.

Source : Galiani, Gertler et Schargrodsky 2005.

Limites de la méthode de la double différence

La méthode de la double différence est généralement moins solide que les méthodes de sélection aléatoire (assignation aléatoire, offre aléatoire et promotion aléatoire). Même si les tendances sont équivalentes entre les deux groupes avant l'intervention, l'estimation peut présenter un biais. En effet, la méthode DD attribue à l'intervention *toute différence de tendances* entre le groupe de traitement et le groupe de comparaison survenant *à partir du lancement de l'intervention*. S'il existe d'autres facteurs influençant la différence de tendances entre les deux groupes, l'estimation sera non valide ou biaisée.

Admettons que vous souhaitiez estimer l'impact du subventionnement de l'achat d'engrais sur la production de riz en mesurant la production des exploitants subventionnés (groupe de traitement) et celle des exploitants non subventionnés (groupe de comparaison), avant et après l'octroi des subventions. Si au cours de la première année, les exploitants subventionnés sont touchés par la sécheresse alors que les exploitants non subventionnés ne le sont pas, la méthode de la double différence produira une estimation incorrecte de l'impact du programme de subventionnement des achats d'engrais. En général, tout facteur affectant uniquement le groupe de traitement et intervenant en même temps que le traitement est susceptible d'invalider ou de biaiser l'estimation d'impact du programme. La méthode de la double différence repose sur *l'hypothèse* que ces facteurs n'existent pas.

Notes

1. Bien que, en théorie, l'assignation aléatoire, la promotion aléatoire et le modèle de discontinuité de la régression ne nécessitent pas de données de référence, en pratique, ces dernières sont essentielles pour confirmer que les caractéristiques du groupe de traitement et du groupe de comparaison sont semblables. Pour cette raison, nous recommandons de collecter des données de base pour toute évaluation. Outre la vérification de la comparabilité des deux groupes, il existe d'autres bonnes raisons de collecter des données de base, même si la méthode utilisée ne l'exige pas. Premièrement, avoir à disposition des caractéristiques (exogènes) de la population avant l'intervention peut permettre de déterminer si le programme a un impact différent au sein de la population éligible en fonction des caractéristiques mesurées avant le programme (analyse d'hétérogénéité). Deuxièmement, les données de base peuvent également permettre d'effectuer une analyse afin d'informer les gestionnaires de programme avant même le début de l'intervention. La collecte des données de base peut par ailleurs servir de pilote à l'échelle pour la collecte de données après l'intervention. Troisièmement, les données de base peuvent servir de « garantie » si l'assignation aléatoire n'est pas mise en œuvre correctement. L'évaluateur peut alors utiliser une combinaison

d'appariement et de double différence pour remédier à d'éventuels problèmes dans la mise en œuvre de l'assignation aléatoire. Enfin, l'existence de données de base peut augmenter la puissance statistique de l'analyse si le nombre d'unités dans le groupe de traitement et de comparaison est limité.

2. Les différences entre les points doivent être interprétées comme des différences verticales sur l'axe vertical de résultat.

Références

DiTella, Rafael et Ernesto Schargrodsky. 2005. « Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack. » *American Economic Review* 94 (1) : 115–33.

Duflo, Esther. 2001. « Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. » *American Economic Review* 91 (4) : 795–813.

Galiani, Sebastian, Paul Gertler et Ernesto Schargrodsky. 2005. « Water for Life: The Impact of the Privatization of Water Services on Child Mortality. » *Journal of Political Economy* 113 (1) : 83–120.

CHAPITRE 7

Appariement

La méthode décrite dans ce chapitre comprend une série de techniques statistiques que nous désignerons collectivement par le terme d'appariement. Les méthodes d'appariement peuvent être appliquées quelles que soient les règles d'assignation de programme, à partir du moment où il existe un groupe qui n'a pas participé au programme. Les méthodes d'appariement utilisent les caractéristiques observées des inscrits et non-inscrits pour générer un groupe de comparaison. Ces méthodes reposent donc sur l'hypothèse très forte qu'il n'y a pas de différence non observée corrélée aux résultats entre le groupe de traitement et le groupe de comparaison. En raison de cette hypothèse très contraignante, les méthodes d'appariement sont généralement plus utiles lorsqu'elles sont combinées à l'une des autres méthodes décrites ci-dessus.

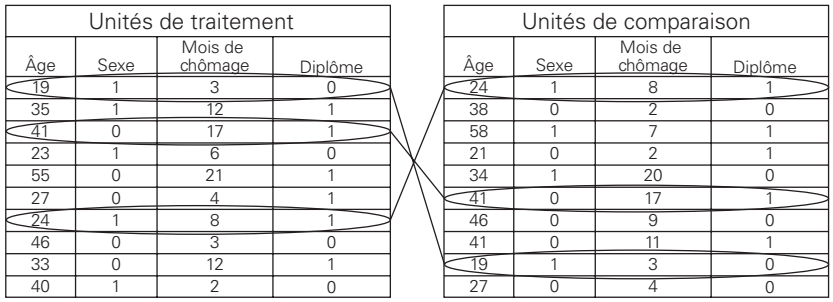
Fondamentalement, l'appariement utilise des techniques statistiques pour produire un groupe de comparaison artificiel en cherchant, pour chaque participant, une observation (ou une série d'observations) du groupe de non-inscrits qui présente des caractéristiques observables les plus semblables possible. Imaginez que vous cherchiez à évaluer l'impact d'un programme et que vous disposiez de données issues d'une enquête démographique et sanitaire à la fois pour les ménages participants et non participants. Le programme que vous cherchez à évaluer n'a pas de règle d'assignation claire (comme l'utilisation de l'assignation aléatoire ou d'un indice d'éligibilité) qui puisse expliquer pourquoi certains ménages participent au programme et d'autres non. Dans ce contexte, les méthodes d'appariement peuvent vous permettre d'identifier les ménages non-inscrits les plus semblables aux ménages inscrits sur la base des caractéristiques observées dans les données. Les ménages non-inscrits « appariés » forment alors le groupe de comparaison servant à estimer le contrefactuel.

Concept clé :
L'appariement consiste à utiliser de grandes bases de données et des techniques statistiques complexes pour générer le meilleur groupe de comparaison artificiel possible pour un groupe de traitement donné.

Pour trouver une unité correspondant au mieux à chaque participant du programme, il est important de définir le plus précisément possible les variables ou *déterminants* expliquant pourquoi chaque individu a décidé de participer au programme ou non. Cette tâche n'est malheureusement pas simple. Si la liste des caractéristiques observées pertinentes est très longue, ou si chaque caractéristique comporte plusieurs valeurs, il peut être difficile de trouver une unité correspondant exactement à chacune des unités du groupe de traitement. Plus le nombre de caractéristiques ou de dimensions des unités à apparier augmente, plus vous risquez d'être confronté à un « problème de dimensionnalité ». Par exemple, si vous n'utilisez que trois caractéristiques pour constituer le groupe de comparaison apparié, par exemple l'âge, le sexe et le lieu de naissance, vous trouverez probablement pour chaque participant des unités correspondantes au sein du groupe des non participants, mais vous courrez le risque de ne pas tenir compte d'autres caractéristiques potentiellement importantes. En revanche, si vous augmentez la liste des variables d'appariement, par exemple, le nombre d'enfants, le nombre d'années d'éducation, l'âge de la mère, l'âge du père, etc., votre base de données risque de ne pas contenir assez d'unités correspondantes pour chaque participant au programme, à moins qu'elle ne contienne un très grand nombre d'observations. La figure 7.1 présente un exemple d'appariement basé sur quatre caractéristiques : l'âge, le genre, le nombre de mois de chômage et le diplôme d'éducation secondaire.

Heureusement, le problème de dimensionnalité peut être évité en utilisant la méthode d'appariement par le score de propension (Rosenbaum et Rubin 1983). Avec cette approche, il n'est pas nécessaire d'apparier chaque participant à un non participant présentant exactement les mêmes caractéristiques observées. Il est suffisant d'estimer la probabilité que chaque participant et non participant s'inscrive

Figure 7.1 Appariement exact sur la base de quatre caractéristiques



au programme sur la base de ses caractéristiques observées. Cette probabilité est appelée le score de propension. Ce score est un chiffre compris entre zéro et un qui résume toutes les caractéristiques observées influençant la participation au programme d'une unité.

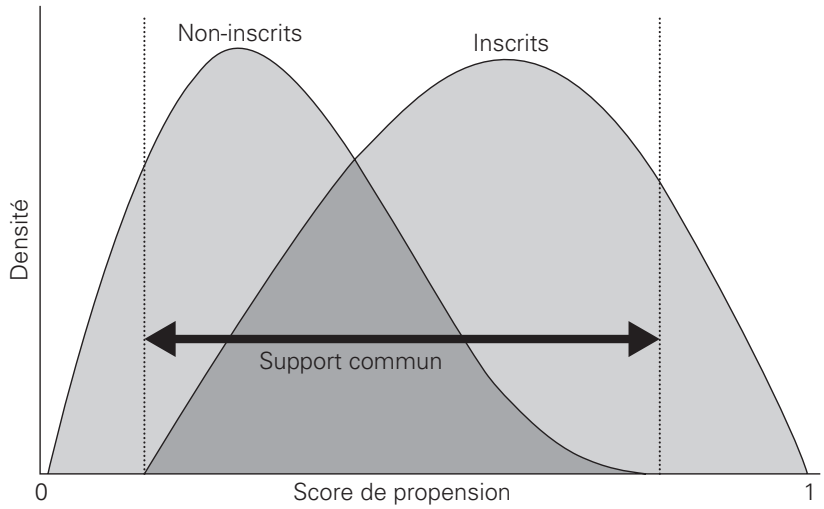
Une fois le score de propension calculé pour toutes les unités, les unités du groupe de traitement peuvent être appariées à celles du groupe de non-inscrits qui affichent le score le plus proche[1]. L'ensemble des « unités les plus proches » forme alors le groupe de comparaison et peut être utilisé pour estimer le contrefactuel. La méthode d'appariement par le score de propension vise à imiter le mécanisme d'assignation aléatoire en choisissant des unités les plus semblables possible aux participants pour constituer le groupe de comparaison. Étant donné que l'appariement par le score de propension n'est pas vraiment une méthode d'assignation aléatoire, mais tente de la répliquer, cette méthode appartient à la catégorie des méthodes quasi expérimentales.

La différence moyenne entre le résultat (Y) des unités soumises au traitement et des unités de comparaison appariées constitue l'impact estimé du programme. En résumé, l'impact du programme est estimé en comparant la moyenne des résultats du groupe de traitement (les participants) à la moyenne des résultats d'un groupe d'unités statistiquement semblables appariées sur la base des caractéristiques observées dans la base de donnée disponible.

Pour que l'appariement par le score de propension débouche sur des estimations valides de l'impact d'un programme, toutes les unités du groupe de traitement doivent pouvoir être appariées à une unité non participante[2]. Cependant, il arrive souvent que pour certaines unités participantes, aucune unité non inscrite ne présente un score de propension similaire. En termes techniques, il s'agit d'un problème de « support commun » entre les scores de propension du groupe de traitement et du groupe des non participants.

La figure 7.2 illustre ce problème de « support commun ». En premier temps, la probabilité que chaque unité de l'échantillon participe au programme est estimée sur la base des caractéristiques observées. Un score de propension est alors attribué à chaque unité. Le score de propension est la probabilité estimée que cette unité participe au programme. La figure illustre la distribution des scores de propension pour les participants et les non participants. Les deux distributions ne se chevauchent pas parfaitement. Pour un score de propension moyen, l'appariement est facile, car les participants et les non participants présentent des caractéristiques similaires. Toutefois, les unités dont le score de propension estimé est proche de zéro ou de un ne peuvent être appariées à un non participant. Intuitivement, les unités qui ont de fortes chances de participer au programme sont tellement différentes des non participants qu'aucune unité correspondante ne peut être trouvée. Le problème de support commun concerne donc les extrémités de la distribution des scores de propension.

Figure 7.2 Appariement par le score de propension et support commun



Jalan et Ravallion (2003a) résument les étapes à suivre pour effectuer un appariement par le score de propension[3]. Premièrement, il faut disposer d'enquêtes représentatives hautement comparables permettant d'identifier à la fois les participants au programme et les non inscrits. Deuxièmement, il faut regrouper les deux échantillons et estimer la probabilité que chaque individu participe au programme sur la base des caractéristiques observées dans les données. Cette étape permet d'obtenir le score de propension. Troisièmement, il faut limiter l'échantillon aux unités pour lesquelles il existe un support commun dans la distribution des scores de propension. Quatrièmement, il s'agit d'identifier, pour chaque unité participante, un sous-groupe d'unités non participantes présentant des scores de propension similaires. Cinquièmement, les résultats des unités participantes peuvent être comparés à ceux des unités non participantes appariées. La différence entre les résultats moyens des deux sous-groupes correspond à l'impact du programme pour l'observation concernée. Sixièmement, la moyenne de ces impacts individuels constitue l'estimation de l'impact moyen du traitement.

En résumé, il convient de retenir deux points importants concernant l'appariement. Premièrement, l'appariement doit être effectué en utilisant les caractéristiques des données de référence collectées avant la mise en place du programme. Deuxièmement, la qualité des résultats obtenus par la méthode d'appariement dépend en grande partie de la qualité des caractéristiques utilisées, et il est donc essentiel de disposer de bases de données très complètes.

Utilisation des techniques d'appariement pour le Programme de subvention de l'assurance maladie (PSAM)

Maintenant que vous comprenez les techniques d'appariement, vous vous demandez comment améliorer les précédentes estimations de l'impact du programme de subvention de l'assurance maladie (PSAM). Vous décidez d'utiliser certaines techniques d'appariement pour sélectionner des groupes de ménages participants et non participants présentant des caractéristiques observées similaires. Vous estimez tout d'abord la probabilité qu'une unité participe au programme sur la base de caractéristiques observées (des « variables explicatives »), telles que l'âge du chef de famille et de son conjoint, leur niveau d'éducation, le genre du chef de famille, l'appartenance du ménage à la population autochtone, etc. Comme l'illustre le tableau 7.1, la probabilité qu'un ménage participe au programme est moindre s'il est plus âgé, s'il a plus d'éducation, s'il est dirigé par une femme ou s'il possède une salle de bains ou une plus grande superficie de terres. En revanche, l'appartenance à la population autochtone, le nombre plus élevé de membres dans le ménage et l'existence d'un sol en terre battue dans le logement sont autant de facteurs qui sont positivement corrélés à la probabilité de participation au

Tableau 7.1 Estimation du score de propension sur la base des caractéristiques observées

Variable dépendante : *Participant* = 1	
Caractéristiques/variables explicatives	**Coefficient**
Âge du chef du ménage (en années)	–0,022**
Âge du conjoint (en années)	–0,017**
Niveau d'éducation du chef du ménage (en années)	–0,059**
Niveau d'éducation du conjoint (en années)	–0,030**
Le chef du ménage est une femme = 1	–0,067
Autochtone = 1	0,345**
Nombre de personnes dans le ménage	0,216**
Sol en terre battue = 1	0,676**
Salle de bains = 1	–0,197**
Hectares de terre	–0,042**
Distance de l'hôpital (en km)	0,001*
Constante	0,664**

Remarque : régression probit. La variable dépendante correspond à 1 si le ménage participe au PSAM et à 0 s'il n'y participe pas. Les coefficients représentent la contribution de chaque caractéristique/variable explicative considérée dans la probabilité qu'un ménage participe au PSAM.

* Seuil de signification de 5 % ; ** Seuil de signification de 1 %.

Tableau 7.2 Cas 7—Impact du PSAM selon la méthode d'appariement (comparaison des moyennes)

	Participants	Non-participants appariés	Différence	Stat. de *t*
Dépenses de santé des ménages	7,8	16,1	–8,3	–13,1

Tableau 7.3 Cas 7— Impact du PSAM selon la méthode d'appariement (analyse de régression)

	Régression linéaire multivariée
Impact estimé sur les dépenses de santé des ménages	–8,3** (0,63)

Remarque : erreurs-types entre parenthèses.

** Seuil de signification de 1 %.

programme. Dans l'ensemble, il semblerait donc que les ménages les plus pauvres et les moins éduqués soient plus susceptibles de participer au programme, ce qui parait encourageant étant donné que le programme cible les ménages pauvres.

Maintenant que vous avez estimé la probabilité que chaque ménage participe au programme (leur score de propension), vous limitez l'échantillon aux ménages participants et non participants que vous pouvez apparier. Pour chaque ménage participant, vous identifiez un sous-groupe de ménages non participants présentant des scores de propension similaires. Le tableau 7.2 compare les résultats moyens pour les ménages participants et les ménages non participants qui leur ont été appariés.

Pour obtenir une estimation d'impact en utilisant la méthode d'appariement, vous devez tout d'abord calculer l'impact individuel pour chaque ménage participant (en le comparant au ménage non participant apparié) puis calculer la moyenne de ces impacts individuels. Selon le tableau 7.3, l'impact estimé grâce à ce procédé correspond à une réduction de 8,3 dollars des dépenses de santé des ménages.

QUESTION 7

A. Quelles sont les hypothèses fondamentales qui sous-tendent le résultat du cas 7 ?
B. Comparez le résultat du cas 7 à celui du cas 3. Pourquoi, selon vous, sont-ils si différents ?
C. Au vu des résultats pour le cas 7, le PSAM doit-il être élargi à l'échelle nationale ?

La méthode d'appariement en pratique

La méthode d'appariement nécessite de grandes bases de données et présente d'autres limites sur le plan statistique, mais elle demeure une méthode relativement polyvalente qui a été utilisée pour évaluer des programmes de développement dans plusieurs contextes. Deux exemples sont décrits en détail dans les encadrés 7.1 et 7.2.

Encadré 7.1 : Programme d'emploi public et revenus en Argentine

Jalan et Ravallion (2003a) utilisent des techniques d'appariement par le score de propension pour évaluer l'impact du programme d'emploi public argentin A Trabajar sur les revenus. En réponse à la crise macroéconomique de 1996-1997 en Argentine, le gouvernement lance rapidement le programme A Trabajar sans avoir recours à des techniques de sélection aléatoire ni collecter des données de base. Par conséquent, Jalan et Ravallion (2003a) utilisent des techniques d'appariement pour évaluer l'impact du programme. Dans ce contexte, les techniques d'appariement servent aussi à analyser si les gains de revenus des ménages varient en fonction du revenu avant l'intervention.

Au milieu de 1997, une enquête est réalisée auprès des participants et des non participants. Afin d'estimer l'impact du programme à l'aide de l'appariement par le score de propension, Jalan et Ravallion mesurent environ 200 caractéristiques des ménages et des communautés. L'estimation des scores de propension montre que les participants au programme sont plus pauvres, plus susceptibles d'être mariés, membres de ménage dont le chef est un homme, plus susceptibles aussi d'être des membres actifs d'associations de quartier.

Après avoir estimé les scores de propension, les auteurs restreignent leur analyse à la région de la distribution des scores de propension où il existe un support commun entre les participants et les non participants. En appariant les participants aux non participants ayant les scores les plus proches et en calculant la moyenne des différences de revenus entre tous ces groupes appariés, les auteurs estiment que le programme entraine une hausse moyenne des revenus équivalant à environ la moitié du salaire du programme public d'emploi. Les chercheurs vérifient la stabilité des résultats en utilisant plusieurs procédures d'appariement. Ils soulignent néanmoins que leurs estimations peuvent être biaisées par certaines caractéristiques non observées. En effet, l'utilisation des méthodes d'appariement ne permet jamais d'exclure la possibilité d'un biais imputable à des variables non observées, ce qui constitue leur principale limite.

Source : Jalan et Ravallion 2003a.

Encadré 7.2 : Eau courante et santé infantile en Inde

Jalan et Ravallion (2003b) utilisent les méthodes d'appariement pour étudier l'impact de l'accès à l'eau courante sur la prévalence et la durée des cas de diarrhée chez les enfants de moins de cinq ans dans les zones rurales en Inde. Les chercheurs évaluent notamment si l'impact de l'extension de l'accès à l'eau dépend des niveaux de revenus ou d'éducation. Cet impact est difficile à mesurer, car il peut également dépendre de comportements parentaux eux aussi susceptibles de réduire l'incidence de la diarrhée, comme par exemple faire bouillir l'eau, assurer une bonne alimentation ou utiliser des sels de réhydratation orale lorsqu'un enfant est malade.

Les chercheurs utilisent des données issues d'une grande enquête d'éducation et de santé menée en 1993-1994 par le National Council of Applied Economic Research auprès de 33 000 ménages ruraux de 16 états indiens. Cette importante base de données permet aux chercheurs de procéder à un appariement par le score de propension à la fois au niveau individuel et au niveau des villages. Ils déterminent le score de propension en estimant la probabilité d'avoir accès à l'eau courante par le biais de la campagne nationale.

L'évaluation conclut que l'accès à l'eau courante entraîne une réduction des cas de diarrhée : la prévalence de diarrhée serait 21 % plus élevée et leur durée 29 % plus longue en l'absence d'eau courante. Toutefois, ces impacts ne sont pas observés dans les groupes à faible revenu, sauf si la femme du foyer a un niveau de scolarité supérieur à l'école primaire. Jalan et Ravallion découvrent que l'impact de l'eau courante sur la santé est plus prononcé dans les ménages où les femmes sont mieux éduquées. Ils concluent qu'il est important de combiner des investissements dans les infrastructures, comme les réseaux d'eau, avec d'autres programmes visant à améliorer l'éducation et à réduire la pauvreté.

Source : Jalan et Ravallion 2003a.

Limites de la méthode d'appariement

Même s'il est possible de procéder à un appariement dans de nombreux contextes et indépendamment des règles d'assignation du programme, cette méthode présente de sérieuses faiblesses.

Premièrement, ces procédures exigent la collecte de grandes bases de données couvrant des échantillons importants. Même si ces bases de données existent, il existe toujours un risque de manque de support commun entre le groupe de traitement et les non participants. Deuxièmement, l'appariement ne peut être effectué que sur la base des caractéristiques observées. Par définition, il n'est pas possible d'intégrer des caractéristiques non observées dans le calcul du score de propension. Ainsi, pour former un groupe de comparaison valide à l'aide de la procédure d'appariement, il faut être sûr qu'il n'existe aucune différence systématique dans les caractéristiques non observées susceptible d'influencer le résultat (Y) entre les participants et des non participants[4].

Il n'est pas possible de *prouver* qu'il n'y a pas de caractéristiques non observées susceptibles d'influer sur la participation et sur les résultats ; il faut donc le *supposer*. Il s'agit en général d'une hypothèse très audacieuse. L'appariement permet de tenir compte des caractéristiques *observées* ; mais ne peut cependant en aucun cas exclure l'existence d'un biais dû aux caractéristiques *non observées*. En résumé, cette hypothèse selon laquelle il n'existe aucun biais de sélection découlant des caractéristiques non observées est très contraignante et ne peut pas être vérifiée, ce qui est problématique.

L'appariement est généralement moins fiable que les autres méthodes d'évaluation déjà évoquées. Par exemple, les méthodes de sélection aléatoire ne reposent pas sur l'hypothèse invérifiable selon laquelle il n'existe pas de variables non observées associées tant à la participation au programme qu'aux résultats. En outre, l'assignation aléatoire ne nécessite pas d'échantillons aussi importants ni de caractéristiques de base aussi nombreuses que la méthode d'appariement.

Dans la pratique, les méthodes d'appariement sont généralement utilisées lorsque la sélection aléatoire, le modèle de discontinuité de la régression et la double différence ne peuvent pas être utilisés. De nombreux évaluateurs utilisent l'appariement a posteriori lorsqu'aucune donnée de base n'est disponible sur le résultat ou les caractéristiques des participants. Ils utilisent une enquête réalisée après le lancement du programme (a posteriori) pour déduire quelles étaient les caractéristiques de la population au départ (par exemple âge, situation de famille), puis ils apparient le groupe de traitement à un groupe de comparaison à partir de ces caractéristiques. Cette approche n'est pas sans risque puisqu'ils peuvent, involontairement, effectuer un appariement sur la base de caractéristiques qui ont été affectées par le programme, ce qui remettrait en question la validité ou l'objectivité de l'estimation.

En revanche, l'appariement à partir des caractéristiques observées dans une enquête de référence collectée avant la mise en œuvre d'un programme peut être très utile s'il est combiné à d'autres techniques comme celle de la double différence, qui tient compte de l'hétérogénéité invariable dans le temps ou non observée. L'appariement est aussi plus utile lorsque la règle d'assignation du programme est connue, auquel cas il peut être effectué sur la base de cette règle (voir chapitre 8).

Les lecteurs auront ici compris qu'il est préférable de concevoir l'évaluation d'impact avant la mise en œuvre d'un programme. Une fois le programme mis en œuvre, s'il n'est pas possible d'influencer la façon dont il est attribué et qu'aucune donnée de base n'a été collectée, il restera peu — voire pas — de possibilités d'évaluation fiables.

Notes

1. Dans la pratique, de nombreuses définitions de ce qui constitue le score de propension le « plus proche » sont utilisées pour réaliser l'appariement. Les unités de comparaison les plus proches peuvent être définies sur la base d'une stratification du score de propension (identification des voisins les plus proches de l'unité de traitement soit en fonction de la distance ou d'un rayon défini) ou en utilisant des techniques non-paramétriques (kernel). Il est conseillé de vérifier la robustesse des résultats obtenus par divers algorithmes d'appariement.

2. La section de ce manuel consacrée à l'appariement se concentre principalement sur l'appariement simple (d'une unité à une autre). D'autres types d'appariement, comme l'appariement d'un à plusieurs ou l'appariement avec ou sans remplacement ne sont pas abordés. Dans tous les cas, l'intuition fondamentale décrite ici s'applique.
3. Rosenbaum (2002) présente une revue détaillée des méthodes d'appariement.
4. Pour les lecteurs versés en économétrie, ceci implique que la participation est indépendante du résultat en conditionnant sur les caractéristiques utilisées pour l'appariement.

Références

Jalan, Jyotsna et Martin Ravallion. 2003a. « Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. » *Journal of Business & Economic Statistics* 21 (1) : 19–30.

———. 2003 b. « Does Piped Water Reduce Diarrhea for Children in Rural India? » *Journal of Econometrics* 112 (1) : 15373.

Rosenbaum, Paul. 2002. *Observational Studies*. 2e éd. Springer Series in Statistics. New York : Springer-Verlag.

Rosenbaum, Paul et Donald Rubin. 1983. « The Central Role of the Propensity Score in Observational Studies of Causal Effects. » *Biometrika* 70 (1) : 41–55.

CHAPITRE 8

Combinaisons de méthodes

Nous avons vu que la plupart des méthodes d'évaluation d'impact ne produisent des estimations valides du contrefactuel que sous certaines hypothèses. Dès lors, le principal risque d'utiliser une méthode donnée est que les hypothèses sur lesquelles elle repose ne soient pas valables et que l'estimation de l'impact du programme soit par conséquent incorrecte. Dans cette section, nous allons résumer ces potentiels problèmes méthodologiques et évoquer les stratégies qui permettent de réduire le risque de biais. Étant donné que ce risque découle principalement de violations des hypothèses sous-jacentes, nous allons nous concentrer sur les approches pour vérifier ces hypothèses.

Il est possible de vérifier la validité des hypothèses qui sous-tendent un certain nombre de méthodes d'évaluation. Pour d'autres méthodes, la véracité des hypothèses ne peut jamais être établie avec certitude, mais différents tests de falsification peuvent contribuer à suggérer que ces hypothèses sont bel et bien valables. Les tests de falsification sont comparables à des tests de résistance : en cas d'échec, il y a de fortes chances que les hypothèses sous-tendant la méthode soient inexactes dans un contexte donné. En revanche, un test réussi n'apporte qu'une indication partielle de la véracité des hypothèses. L'encadré 8.1 présente une liste de tests de vérification et de falsification qui peuvent être utilisés pour évaluer la pertinence d'une méthode d'évaluation dans un contexte particulier. La liste contient certaines questions pratiques dont les réponses peuvent être obtenues en analysant les données de l'enquête de base.

Encadré 8.1 : Liste des tests de vérification et de falsification

Assignation aléatoire

L'assignation aléatoire est considérée comme la méthode la plus rigoureuse pour évaluer le contrefactuel. Il s'agit de « l'étalon-or » de l'évaluation d'impact. De simples tests sont toutefois nécessaires pour jauger de la validité de cette stratégie d'évaluation dans un contexte donné.

- Les caractéristiques sont-elles équilibrées dans les données de référence ? Comparez les caractéristiques du groupe de traitement à celles du groupe de comparaison en utilisant les données de base[a].
- Les unités ont-elles totalement adhéré au résultat de l'assignation aléatoire ? Vérifiez si toutes les unités éligibles ont bien bénéficié du programme et qu'aucune unité non éligible n'en a bénéficié. Si l'adhérence n'est pas totale, utilisez la méthode de l'offre aléatoire.
- Le nombre d'unités dans le groupe de traitement et dans le groupe de comparaison est-il assez important ? Si ce n'est pas le cas, considérez combiner l'assignation aléatoire et la double différence.

Offre aléatoire

Si l'adhérence n'est pas totale, l'assignation aléatoire revient à l'offre aléatoire.

- Les caractéristiques sont-elles équilibrées dans les données de référence ? Comparez les caractéristiques des unités auxquelles le programme a été offert à celles des unités auxquelles il n'a pas été offert en utilisant les données de base.

Promotion aléatoire

La promotion aléatoire fournit une estimation valide du contrefactuel si la campagne de promotion augmente substantiellement la participation au programme sans influencer directement le résultat à l'étude.

- Les caractéristiques des unités recevant la campagne de promotion et celles ne la recevant pas sont-elles équilibrées dans l'enquête de référence ? Comparez les caractéristiques des deux groupes en utilisant les données de base.
- La campagne de promotion augmente-t-elle significativement la participation au programme ? Elle le devrait. Comparez les taux de participation au programme entre le groupe ayant reçu une promotion et le groupe ne l'ayant pas reçu.
- La campagne de promotion a-t-elle un effet direct sur les résultats ? Elle ne devrait pas en avoir. Ceci ne peut généralement pas être testé directement et il faut donc se fier aux théories existantes et au bon sens.

Modèle de discontinuité de la régression

Pour pouvoir utiliser le modèle de discontinuité de la régression, il faut que l'indice d'éligibilité soit continu aux alentours du seuil d'éligibilité et que les unités proches du seuil soient comparables.

- L'indice est-il continu aux alentours du seuil d'éligibilité dans les données de référence ?
- L'adhérence au seuil d'éligibilité est-elle totale ? Vérifiez si toutes les unités éligibles ont bénéficié du programme et qu'aucune unité non éligible n'en a bénéficié. Si vous découvrez que l'adhérence au seuil d'éligibilité n'est pas totale, combinez le modèle de discontinuité de la régression avec des techniques plus sophistiquées pour corriger cette « discontinuité floue » ("fuzzy discontinuity" en anglais)[b].

Double différence (DD)

La méthode de la double différence part du principe que les tendances du résultat sont similaires pour le groupe de traitement et le groupe de comparaison avant l'intervention, et que les seuls facteurs à l'origine de changements du résultat entre les deux groupes sont constants dans le temps.

- Les résultats du groupe de traitement et du groupe de comparaison auraient-ils évolué en parallèle en l'absence du programme ? Il est possible de répondre à cette question en utili-

(suite)

Encadré 8.1 *suite*

sant plusieurs tests de falsification : 1) Les résultats du groupe de traitement et du groupe de comparaison évoluent-ils en parallèle avant l'intervention ? Si deux rondes de données sont disponibles avant le lancement du programme, vérifiez si les tendances des deux groupes divergent. 2) Qu'en est-il des faux résultats qui ne devraient pas être affectés par le programme ? Évoluent-ils en parallèle avant et après l'intervention pour le groupe de traitement et pour le groupe de comparaison ?

- Effectuez l'analyse de la double différence en utilisant plusieurs groupes de comparaison plausibles. Obtenez-vous des estimations similaires de l'impact du programme ?
- Effectuez l'analyse de la double différence en utilisant le groupe de traitement et le groupe de comparaison de votre choix et un faux résultat qui ne devrait pas être affecté par le programme. Vous devriez obtenir un impact nul du programme sur ce résultat.
- Effectuez l'analyse de la double différence en utilisant la variable de résultat de votre choix et deux groupes qui n'ont pas été affectés par le programme. Vous devriez obtenir un impact nul du programme.

Appariement

L'appariement repose sur l'hypothèse selon laquelle les unités participantes et les unités non participantes sont similaires au niveau des variables non observées qui pourraient affecter leur probabilité de participation au programme et le résultat (*Y*).

- La participation au programme est-elle déterminée par des variables non observables ? Ceci ne peut généralement pas être directement vérifié et il faut donc se fier aux théories existantes et au bon sens.
- Les caractéristiques observées des groupes appariés sont-elles bien équilibrées ? Comparez les caractéristiques observées de chaque unité du groupe de traitement et de son unité appariée du groupe de comparaison.
- Pouvez-vous apparier chaque unité de traitement avec une unité de comparaison ? Vérifiez qu'il existe un support commun suffisant dans la distribution des scores de propension. Un support commun limité indique que les participants et les non participants sont très différents, suggérant que l'appariement n'est peut-être pas la méthode la plus pertinente.

a. Comme nous l'avons déjà indiqué, pour des raisons statistiques, il n'est pas nécessaire que toutes les caractéristiques observées dans le groupe de traitement et dans le groupe de comparaison soient similaires pour que l'assignation aléatoire puisse être considérée comme efficace. Même si les caractéristiques des deux groupes sont entièrement similaires, on peut s'attendre à ce que 5 % des caractéristiques présentent une différence statistiquement significative en utilisant un niveau de confiance de 95 % pour le test.

b. Nous n'aborderons pas cette technique dans ce manuel, mais elle consiste à combiner le modèle de discontinuité de la régression avec une variable instrumentale. Il s'agit d'utiliser le seuil d'éligibilité comme variable instrumentale pour la participation effective des unités au programme dans la première étape d'une méthode des moindres carrés à deux étapes.

Combinaisons de méthodes

Même si toutes les méthodes d'évaluation comportent des risques de biais, il est parfois possible de les limiter en combinant plusieurs méthodes. La combinaison de plusieurs méthodes permet en effet de compenser les limites d'une méthode donnée et ainsi de renforcer la solidité de l'estimation du contrefactuel.

La *double différence appariée* (DD appariée) est un exemple de combinaison de méthodes. Comme mentionné ci-dessus, le simple appariement par le score de propension ne tient pas compte des caractéristiques non observées qui peuvent expliquer pourquoi un groupe a choisi de participer à un programme et qui sont également susceptibles d'affecter les résultats. En revanche, la combinaison de l'appariement et de la double différence permet au moins de contrôler pour les caractéristiques non observées qui sont constantes dans le temps pour les deux groupes. Elle est appliquée comme suit :

- Premièrement, effectuez l'appariement sur la base des caractéristiques observées dans les données de base (voir chapitre 7).
- Deuxièmement, appliquez la méthode de la double différence afin d'estimer un contrefactuel pour le changement du résultat pour chaque sous-groupe d'unités appariées.
- Troisièmement, calculez la moyenne de ces doubles différences pour tous les sous-groupes.

L'encadré 8.2 fournit un exemple concret d'évaluation basée sur la méthode de la double différence appariée.

Il est également possible de combiner le *modèle de discontinuité de la régression et la double différence*. Souvenez-vous que le modèle de discontinuité de la régression part du principe que les unités aux alentours du seuil d'éligibilité sont très similaires. Dans la mesure où des différences demeurent entre les unités des deux côtés du seuil d'éligibilité, l'utilisation de la double différence permet de contrôler pour les différences dans les caractéristiques non observées constantes dans le temps. La combinaison du modèle de discontinuité de la régression et de la double différence peut être appliquée en calculant la double différence du résultat pour les unités de part et d'autre du seuil d'éligibilité.

Adhérence non totale

Une différence entre le traitement prévu et le traitement effectif pour certaines unités signifie que l'adhérence au programme n'est pas totale. Nous avons abordé ce point dans le contexte de l'assignation aléatoire, mais il s'agit d'un problème qui peut concerner la plupart des méthodes d'évaluation d'impact. Avant de pouvoir interpréter l'impact estimé à l'aide d'une méthode, quelle qu'elle soit, vous devez déterminer si l'adhérence au programme est totale ou pas.

L'adhérence n'est pas totale dans deux cas distincts : 1) certaines unités ciblées peuvent ne pas avoir participé au traitement et 2) certaines unités de comparaison peuvent avoir participé au traitement. L'adhérence peut ne pas être totale pour plusieurs raisons :

- Tous les participants ciblés par le programme n'y participent pas. Parfois, les unités auxquelles le programme est proposé choisissent de ne pas y participer.
- Le programme n'est pas offert à certains participants ciblés en raison d'une erreur administrative ou de mise en œuvre.

Encadré 8.2 : Double différence appariée

Sols en ciment, santé infantile et bonheur maternel au Mexique

Le programme Piso Firme au Mexique propose d'installer jusqu'à 50 mètres carrés de sol en ciment dans les logements dont le sol est en terre battue. Piso Firme a été lancé comme un programme local dans l'État de Coahuila avant d'être adopté à l'échelle nationale. Cattaneo et al. (2009) profitent de la variation géographique dans la distribution du programme pour évaluer l'impact de l'amélioration des logements sur la santé et les conditions de vie.

Les chercheurs utilisent la méthode de la double différence combinée à celle de l'appariement pour comparer les ménages de Coahuila à des foyers similaires dans l'État voisin de Durango où, à l'époque de la réalisation de l'enquête, le projet n'avait pas encore été mis en œuvre. Pour améliorer la comparabilité entre le groupe de traitement et le groupe de comparaison, les chercheurs limitent leur échantillon aux ménages des villes voisines se situant de chaque côté de la frontière entre les deux États. Ils prélèvent leurs échantillons dans les quartiers des deux villes présentant des caractéristiques similaires avant l'intervention au moment du recensement de 2002.

En utilisant l'offre de sol en ciment comme une variable instrumentale pour la possession effective d'un sol en ciment, les chercheurs estiment le traitement sur les traités à partir des estimations de l'intention de traiter et découvrent que le programme entraine une réduction de 18,2 % de la présence de parasites, de 12,4 % de la prévalence de la diarrhée et de 19,4 % de la prévalence d'anémie. Ils sont par ailleurs en mesure d'utiliser la variation de la surface totale au sol couverte par du ciment pour prédire qu'un remplacement intégral des sols en terre battue par des sols en ciment dans les logements entraînerait une réduction de 78 % des infections parasitaires, de 49 % des cas de diarrhée et de 81 % des cas d'anémie tout en augmentant le développement cognitif de 36 % à 96 %. Les auteurs collectent également des données sur les conditions de vie des adultes et découvrent que les sols en ciment rendent aussi les mères plus heureuses, ce qui se manifeste par une augmentation de 59 % de la satisfaction à l'égard du logement, de 69 % de satisfaction à l'égard de la qualité de vie, et par une baisse de 52 % du score obtenu sur une échelle d'évaluation de la dépression et de 45 % du score obtenu sur l'échelle d'évaluation du stress.

Cattaneo et al. (2009) concluent leur rapport en montrant que le programme Piso Firme a eu un impact absolu plus marqué sur le développement cognitif des enfants pour un coût inférieur à celui du programme national mexicain de transferts monétaires conditionnels (Oportunidades/Progresa) et d'autres programmes comparables de suppléments alimentaires ou de stimulation cognitive pour les enfants en bas âge. Les sols en ciment ont également un effet préventif sur les infections parasitaires plus efficace que les traitements vermifuges habituels. Les auteurs indiquent que les programmes visant à remplacer les sols en terre battue par des sols en ciment constituaient un moyen abordable d'améliorer la santé infantile dans des contextes similaires.

Source : Cattaneo et al. 2009.

- Le programme a été proposé par erreur à des unités du groupe de comparaison, qui y participent.
- Certaines unités du groupe de comparaison parviennent à participer au programme bien qu'il ne leur soit pas proposé, ce qui est parfois caractérisé de débordement ou de « contamination » du groupe de comparaison. Si les effets de débordements touchent une grande partie du groupe de comparaison, il peut devenir impossible d'obtenir une estimation objective du contrefactuel.
- L'assignation du programme repose sur un score continu, mais le seuil d'éligibilité n'est pas strictement respecté.
- Une migration sélective s'opère en raison du programme. Par exemple, la méthode de la double différence peut être utilisée pour comparer les résultats des municipalités traitées et non traitées, mais certains particuliers peuvent choisir de se déplacer d'une municipalité à l'autre s'ils n'apprécient pas qu'elle reçoive ou non le programme.

En général, si l'adhérence n'est pas totale, les méthodes d'évaluation d'impact standard produisent des estimations de l'intention de traiter. Les estimations du traitement sur les traités peuvent toutefois être calculées à partir des estimations de l'intention de traiter en utilisant une variable instrumentale.

Au chapitre 4, nous avons présenté l'intuition pour faire face au manque d'adhérence totale dans le contexte de l'assignation aléatoire. En ajustant le pourcentage des adhérents dans l'échantillon d'évaluation, nous sommes en mesure de mesurer l'impact du traitement sur les traités à partir de l'estimation de l'intention de traiter. Cette technique peut s'appliquer à d'autres méthodes en utilisant l'approche plus générale de variable instrumentale. La variable instrumentale est une variable qui permet de résoudre ou de corriger le manque d'adhérence totale. Dans le cas de l'offre aléatoire, nous utilisons une variable 0/1 (ou variable binaire) dont la valeur est un si l'unité était initialement incluse dans le groupe de traitement et zéro si elle était initialement intégrée au groupe de comparaison. Au moment de l'analyse, la variable instrumentale est souvent utilisée dans le contexte d'une *régression en deux étapes* qui permet de déterminer l'impact du traitement sur les adhérents.

La logique de la technique de variable instrumentale peut être appliquée à d'autres méthodes d'évaluation :

- Dans le contexte du modèle de discontinuité de la régression, la variable instrumentale à utiliser est une variable 0/1 qui indique où se situe une unité par rapport au seuil d'éligibilité.
- Dans le contexte de la double différence et de la migration sélective, la localisation d'un individu avant l'annonce du programme peut servir de variable instrumentale pour la localisation de l'individu après le lancement du programme.

Bien qu'il soit possible de « corriger » un manque d'adhérence totale en utilisant des variables instrumentales, il convient de souligner deux points :

1. D'un point de vue technique, il n'est pas souhaitable qu'une large proportion du groupe de comparaison participe au programme. Les évaluateurs et les décideurs impliqués dans l'évaluation d'impact doivent travailler ensemble pour faire en sorte de limiter cette proportion.
2. La méthode à variable instrumentale n'est valide que dans certaines circonstances et ne constitue pas une solution universelle.

Effets de diffusion

Même si le groupe de comparaison ne participe pas directement au programme, il peut bénéficier indirectement d'un effet de diffusion (ou de débordement) découlant du groupe de traitement. Kremer et Miguel (2004) examinent l'impact de la distribution de médicaments vermifuges aux enfants dans les écoles kenyanes et présentent un exemple intéressant de ce phénomène (encadré 8.3). Les vers intestinaux sont des parasites qui peuvent être transmis d'une personne à l'autre par contact avec des matières fécales contaminées. Lorsqu'un enfant prend des médicaments vermifuges, son degré d'infestation par les vers diminue. Les personnes vivant dans le même environnement que cet enfant sont à leur tour en contact avec moins de vers. Ainsi, dans l'exemple kenyan, la distribution de vermifuges aux enfants d'une école bénéficie non seulement aux enfants de cette école (un effet direct), mais également à ceux des écoles voisines (un effet indirect).

Comme le montre la figure 8.1, la distribution de vermifuges aux écoles du groupe A permet de réduire le nombre de vers chez les enfants des écoles du groupe B ne participant pas au programme, mais se situant à proximité des écoles du groupe A. En revanche, les écoles non participantes éloignées des écoles du groupe A (écoles du groupe C) ne sont pas touchées par les effets de diffusion, car la distribution de médicaments au groupe A n'a pas d'effet indirect sur les vers touchant le groupe C. Kremer et Miguel (2004) concluent que le traitement vermifuge réduit fortement le taux d'absentéisme non seulement dans les écoles participant au programme (comparaison entre le groupe A et le groupe C), mais également dans les écoles non participantes voisines (comparaison entre le groupe B et le groupe C).

Quand des effets de débordements sont possibles, il est important que l'évaluateur vérifie qu'ils n'affectent pas l'ensemble du groupe de comparaison. Pour autant que suffisamment d'unités de comparaison ne soient pas affectées par les effets de diffusion (le groupe C dans l'exemple du traitement vermifuge), vous pourrez estimer l'impact du programme en comparant les résultats des unités du groupe de traitement et ceux des unités du groupe de comparaison non affecté. L'inconvénient est que l'évaluation ne pourra pas permettre de généraliser l'estimation des effets du traitement à l'ensemble de la population. Lors de la conception de l'évaluation, si vous pensez qu'un programme engendrera des effets de débordements, vous pouvez ajuster la méthode d'évaluation afin de produire de meilleurs résultats. Premièrement, l'évaluation doit pouvoir compter sur un

Encadré 8.3 : Programme avec effets de diffusion

Traitement vermifuge, effets externes et éducation au Kenya

Le projet de traitement vermifuge dans les écoles primaires de Busia au Kenya a été mis en œuvre par l'organisation néerlandaise à but non lucratif Child Support Africa en coopération avec le ministère de la Santé. Il est conçu pour étudier divers aspects de la prévention et du traitement vermifuges. Le projet couvre initialement 75 écoles, soit plus de 30 000 élèves âgés de six à 18 ans. Les écoles bénéficient de distribution de médicaments vermifuges conformément aux recommandations de l'Organisation mondiale de la santé ainsi que d'une formation préventive comprenant des présentations sur le thème de la santé, de diagrammes muraux et de cours destinés aux enseignants.

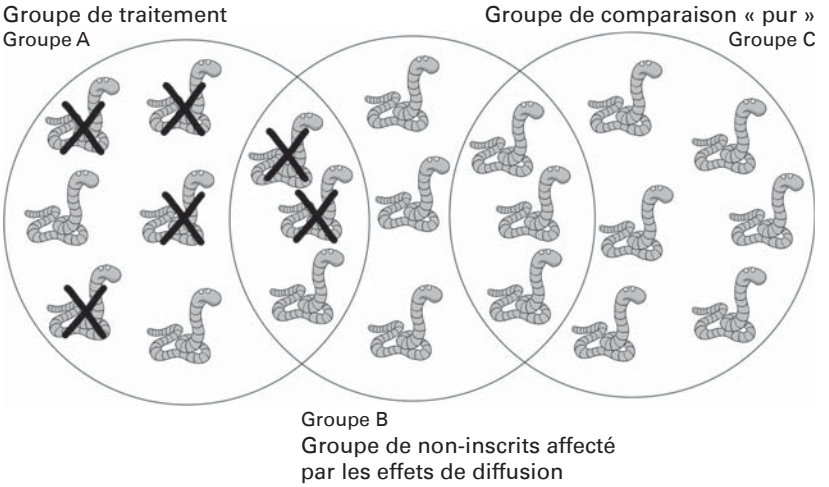
En raison de contraintes administratives et financières, le programme a été graduellement déployé par ordre alphabétique, le premier groupe de 25 écoles en bénéficiant dès 1998, le deuxième groupe en 1999 et le troisième groupe en 2001. En utilisant cette assignation aléatoire au niveau des écoles, Kremer et Miguel (2004) sont en mesure d'estimer l'impact du traitement vermifuge sur une école et de déterminer les effets de diffusion entre les écoles en utilisant la variation exogène de la proximité des écoles de comparaison au groupe de traitement. Malgré une adhérence à l'assignation aléatoire relativement élevée (75 % des élèves ciblés par le traitement vermifuge reçoivent des médicaments, contre seulement un faible pourcentage des unités du groupe de comparaison), les chercheurs sont également en mesure d'exploiter le manque d'adhérence totale pour déterminer les externalités ou effets de débordements à l'intérieur même des écoles traitées. Kremer et Miguel (2004) concluent que les externalités à l'intérieur même des écoles traitées entrainent une réduction de 12 points d'infections modérées à élevées tandis que l'effet direct supplémentaire lié à la prise de médicament vermifuge correspond à 14 points supplémentaires. Par ailleurs, en termes d'externalités entre les écoles, ils estiment à 26 points la baisse des infections modérées à élevées par tranche de 1 000 élèves inscrits dans une école du groupe de traitement. Ces effets sur la santé sont également accompagnés d'une hausse du taux de fréquentation de l'école d'au moins 7 % et d'une réduction de l'absentéisme d'au moins 25 %. Aucun impact significatif sur les résultats aux examens n'est relevé.

Au vu du faible coût du traitement vermifuge et de ses effets relativement importants sur la santé et l'éducation, les chercheurs concluent que le traitement vermifuge est un moyen relativement rentable d'améliorer les taux de scolarisation. L'étude indique également que les maladies tropicales comme les vers peuvent jouer un rôle important sur l'éducation et renforce la théorie selon laquelle la forte charge de morbidité dont souffre l'Afrique contribue peut-être à son faible niveau de revenu. Les auteurs du rapport recommandent donc le renforcement des subventions publiques pour les traitements médicaux présentant des effets de diffusion similaires dans les pays en développement.

Source : Kremer et Miguel 2004.

groupe de comparaison « pur » qui permette de généraliser l'estimation de l'impact du programme. Deuxièmement, la méthodologie peut rendre possible l'estimation de l'ampleur des effets de débordements si elle génère un groupe de comparaison qui bénéficie seulement de ces effets indirects. Les débordements sont souvent intéressants au niveau politique, car ils constituent des impacts indirects des programmes.

Figure 8.1 Effets de diffusion



La figure 8.1 montre qu'il est possible d'estimer à la fois l'impact d'un programme et ses éventuels effets de diffusion. Les médicaments sont distribués au groupe A. Les effets du traitement se propagent au groupe B. Le groupe C est plus éloigné et ne bénéficie donc pas des effets de diffusion. Ce scénario peut être obtenu par l'assignation aléatoire du traitement entre deux unités rapprochées et à une unité similaire plus éloignée. Dans ce cadre simple, l'impact du programme peut être estimé en comparant les résultats du groupe A à ceux du groupe C, et les effets de diffusion peuvent être estimés en comparant les résultats du groupe B à ceux du groupe C.

Considérations supplémentaires

Outre le manque d'adhérence totale et les effets de diffusion, d'autres facteurs doivent être pris en compte au moment de l'élaboration d'une évaluation d'impact. Ces facteurs sont communs à la plupart des méthodologies que nous avons abordées et ils sont généralement plus difficiles à atténuer[1].

Au moment de la planification d'une évaluation, il convient de déterminer le meilleur moment pour collecter les données. S'il faut attendre longtemps avant qu'un programme exerce un impact sur les résultats, collecter les données trop tôt

peut impliquer une estimation d'impact nulle (voir par exemple King et Behrman 2009). Au contraire, si l'enquête de suivi est réalisée trop tard, vous ne serez pas en mesure d'évaluer les effets du programme à temps pour informer les décideurs. Si vous souhaitez évaluer à la fois l'impact à court terme et à long terme du programme, vous devrez collecter plusieurs rondes de données de suivi après l'intervention. Le chapitre 10 contient des informations complémentaires pour déterminer le calendrier de l'évaluation.

Si vous souhaitez estimer l'impact d'un programme sur un groupe entier, vous risquez de passer à côté de certaines variations des impacts entre les différents bénéficiaires du traitement. La plupart des méthodes d'évaluation partent du principe qu'un programme affecte les résultats de manière simple et linéaire pour toutes les unités de la population étudiée. Des problèmes peuvent toutefois survenir lorsque l'ampleur de la réaction dépend de façon non linéaire de l'ampleur de l'intervention ou lorsqu'un groupe recevant un traitement de forte intensité est comparé à un groupe recevant un traitement de faible intensité. Si vous pensez que différents sous-groupes sont susceptibles de réagir différemment au programme, vous pouvez envisager de former des échantillons séparés pour chaque sous-groupe. Admettons que vous cherchiez à connaître l'impact d'un programme de repas scolaires sur les filles, mais qu'elles ne représentent que 10 % des élèves. Dans ce cas, il est possible que même un large échantillon d'élèves ne contienne pas un nombre suffisant de filles pour vous permettre d'estimer l'impact du programme sur celles-ci. Il vous faudra donc stratifier votre échantillon en fonction du genre et inclure un nombre suffisant de filles dans l'échantillon final pour vous permettre d'identifier un impact donné.

Lorsque vous réalisez une évaluation d'impact, il est possible que vous provoquiez involontairement des changements de comportements au sein de la population à l'étude, ce qui peut limiter la validité externe des résultats de votre évaluation. Par exemple, l'effet Hawthorne se produit lorsque le fait même d'être observées provoque un changement de comportement chez les unités (Levitt et List 2009). L'effet John Henry se produit lorsque les unités de comparaison font des efforts supplémentaires pour compenser l'absence de traitement. L'anticipation peut entraîner un autre type de comportement involontaire. Dans le cadre d'un déploiement aléatoire d'un programme, les unités du groupe de comparaison peuvent s'attendre à bénéficier du programme à l'avenir et donc commencer à changer de comportement avant même que le programme ne leur parvienne. Si vous avez des raisons de penser que ces comportements involontaires existent, la création de groupes de comparaison supplémentaires qui ne sont en aucune façon affectés par l'intervention peut être une option qui vous permet de contrôler pour ces comportements, ou même de mesurer explicitement leur amplitude.

Un plan de rechange pour votre évaluation

Même si on est armé de la meilleure méthode d'évaluation d'impact et qu'on est animé des meilleures intentions, les choses ne se passent pas toujours comme prévu. Dans le cadre d'un récent programme de formation professionnelle, l'organisme responsable de la mise en œuvre du programme pensait que beaucoup de candidats allaient s'inscrire et avait projeté de sélectionner les participants de manière aléatoire à partir du groupe de candidats. En raison d'un taux de chômage élevé au sein de la population ciblée, l'organisme pensait que le nombre de candidats au programme de formation professionnelle serait nettement supérieur au nombre de places disponibles. Malheureusement, la campagne de promotion du programme a été moins efficace qu'on l'a espéré et, au final, le nombre de candidats s'est avéré légèrement inférieur au nombre de places disponibles. En l'absence d'un nombre suffisant de candidats pour pouvoir former un groupe de comparaison et faute d'un plan de rechange, le projet initial d'évaluation du programme a dû être abandonné. Ce type de situation est fréquent, tout comme les changements inattendus de contexte opérationnel ou politique. Il est donc utile d'avoir un plan de rechange au cas où la méthodologie choisie initialement ne peut pas être appliquée. La partie 3 du présent manuel aborde plus en détail les aspects opérationnels et politiques de l'évaluation.

Planifier plusieurs méthodes d'évaluation d'impact est également une bonne pratique méthodologique. Si vous avez des doutes qu'une méthode souffre d'un éventuel biais, l'utilisation d'une méthode complémentaire permet de vérifier les résultats. Lorsqu'un programme fait l'objet d'un déploiement aléatoire (voir chapitre 10), le groupe de comparaison est au final intégré dans le programme, ce qui limite la durée pendant laquelle le programme peut être évalué. Toutefois, si la méthode de promotion aléatoire est appliquée en plus de l'assignation aléatoire, vous disposerez d'un groupe de comparaison pour toute la durée du programme. Avant l'incorporation du dernier groupe au programme, vous aurez deux autres groupes de comparaison (ceux obtenus par l'assignation aléatoire et par la promotion aléatoire), bien qu'à long terme il ne vous restera plus que le groupe de comparaison généré par la promotion aléatoire.

Note

1. Le chapitre 3 aborde d'autres facteurs limitant la validité externe liés aux biais d'échantillonnage ou à différents niveaux d'attrition pour le groupe de traitement et le groupe de comparaison.

Références

Cattaneo, Matias, Sebastian Galiani, Paul Gertler, Sebastian Martinez et Rocio Titiunik. 2009. « Housing, Health and Happiness. » *American Economic Journal : Economic Policy* 1 (1) : 75–105.

King, Elizabeth M. et Jere R. Behrman. 2009. « Timing and Duration of Exposure in Evaluations of Social Programs. » *World Bank Research Observer* 24 (1) : 55–82.

Kremer, Michael et Edward Miguel. 2004. « Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. » *Econometrica* 72 (1) : 159–217.

Levitt, Steven D. et John A. List. 2009. « Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. » NBER Working Paper 15016, National Bureau of Economic Research, Cambridge, MA.

CHAPITRE 9

Évaluation de programmes à multiples facettes

Jusqu'à présent, nous nous sommes concentrés sur des programmes qui ne comprennent qu'un seul type de traitement. Dans la pratique, de nombreuses questions politiques pertinentes se posent concernant des programmes à multiples facettes, par exemple des programmes qui combinent plusieurs types de traitement[1]. Les décideurs peuvent non seulement chercher à savoir si un programme est efficace, mais aussi s'il fonctionne mieux ou est plus rentable qu'un autre. Par exemple, en vue d'accroître le taux de scolarisation, est-il plus efficace de mettre en œuvre des interventions qui influencent la demande (comme les transferts monétaires aux familles) ou l'offre (une meilleure rémunération des enseignants) ? La mise en place conjointe de ces deux types d'interventions serait-elle plus efficace que chaque intervention réalisée séparément ? Autrement dit, sont-elles complémentaires ? D'autre part, si la rentabilité des programmes est une priorité, vous pouvez vous demander quel est le niveau de services optimal que le programme doit fournir. Par exemple, quelle est la durée optimale d'un programme de formation professionnelle ? Un programme de six mois permet-il à un plus grand nombre de participants de trouver un emploi qu'un programme de trois mois ? Le cas échéant, la différence de résultats est-elle suffisante pour justifier la mobilisation des ressources supplémentaires pour mettre en œuvre un programme de six mois ?

Au-delà de la simple estimation de l'impact d'une intervention sur le résultat à l'étude, les évaluations d'impact peuvent permettre de répondre à des questions plus générales :

- *Quel est l'impact d'un traitement comparé à l'impact d'un autre traitement ?* Par exemple, quel est l'impact sur le développement cognitif des enfants d'un programme d'éducation parental en comparaison à l'impact d'un programme de nutrition ?

- *L'impact cumulé de deux traitements est-il plus important que la somme des impacts de chaque traitement pris séparément ?* Par exemple, l'impact global du programme d'éducation parental et du programme de nutrition est-il plus important, équivalent ou moins important que la somme des impacts des deux interventions pris séparément ?
- *Quel est l'impact supplémentaire d'un traitement à forte intensité comparé à un traitement à faible intensité ?* Par exemple, quel est l'impact sur le développement cognitif des enfants en retard de croissance de la visite à domicile d'un travailleur social toutes les deux semaines en comparaison à une seule visite mensuelle ?

Ce chapitre illustre comment élaborer des évaluations d'impact pour plusieurs types de programmes à multiples facettes : ceux qui offrent un traitement qui peut avoir une intensité variable, et ceux qui contiennent plusieurs types de traitements. Nous abordons dans un premier temps les méthodes d'élaboration d'évaluation d'impact de programme avec plusieurs niveaux de bénéfices potentiels, puis nous étudierons comment distinguer les différents types d'impact d'un programme comportant plusieurs traitements. Les exemples donnés reposent sur l'utilisation du mécanisme d'assignation aléatoire, mais peuvent être également appliqués à d'autres méthodes.

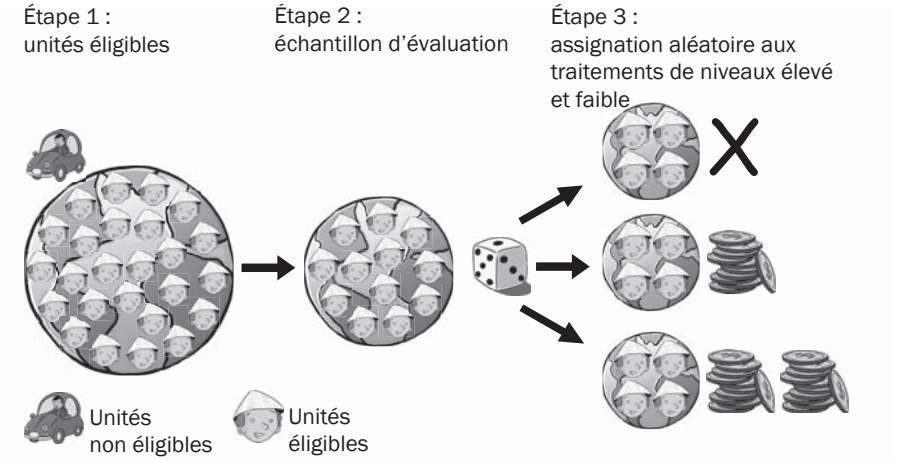
Évaluation de programmes à différents niveaux de traitement

Il est relativement simple d'élaborer une évaluation d'impact pour un programme qui présente différents niveaux de traitement. Imaginez que vous cherchez à évaluer l'impact d'un programme présentant deux intensités possibles du traitement : un niveau élevé (par exemple, des visites toutes les deux semaines) et un niveau faible (par exemple, des visites mensuelles). Vous voulez évaluer l'impact des deux options pour savoir dans quelle mesure la fréquence des visites influe sur les résultats. Pour cela, vous pouvez désigner par tirage au sort les bénéficiaires du traitement de niveau d'intensité élevé, les bénéficiaires du traitement de faible niveau d'intensité et les membres du groupe de comparaison. La figure 9.1 illustre ce processus.

Comme pour une assignation aléatoire ordinaire, la première étape consiste à définir les unités éligibles au programme. La deuxième étape consiste à sélectionner un échantillon d'unités pour l'évaluation, à savoir l'échantillon d'évaluation. Une fois l'échantillon d'évaluation créé, la troisième étape consiste à répartir les unités de façon aléatoire entre le groupe bénéficiant du traitement à intensité élevé, le groupe bénéficiant du traitement à faible intensité, et le groupe de comparaison. L'assignation aléatoire des unités à différents niveaux de traitement permet d'obtenir trois groupes distincts :

- Le groupe A est le groupe de comparaison.

Figure 9.1 Étapes de l'assignation aléatoire à deux niveaux de traitement



- Le groupe B reçoit le traitement de faible niveau d'intensité.
- Le groupe C reçoit le traitement de niveau d'intensité élevé.

Si elle est correctement effectuée, l'assignation aléatoire permet de créer trois groupes similaires. Vous pouvez donc estimer l'impact du traitement de niveau élevé en comparant le résultat moyen du groupe C à celui du groupe A. Vous pouvez également estimer l'impact du traitement de faible intensité en comparant le résultat moyen du groupe B à celui du groupe A. Enfin, vous pouvez déterminer si le traitement de niveau élevé a un impact plus important que le traitement de faible niveau en comparant les résultats moyens des groupes B et C.

L'estimation de l'impact d'un programme comportant plus de deux niveaux de traitement suit la même logique. S'il existe trois niveaux d'intensité de traitement, le processus d'assignation aléatoire donne lieu à la création de trois groupes de traitement en plus du groupe de comparaison. En général, pour *n* niveaux de traitement, vous aurez *n* groupes de traitement plus un groupe de comparaison.

Lorsqu'il n'est pas possible de procéder à une assignation aléatoire, d'autres méthodes d'évaluation peuvent être appliquées. Toutes les méthodes d'évaluation décrites jusqu'à présent permettent d'analyser l'impact relatif de différents niveaux de traitement. Imaginons par exemple que vous souhaitiez évaluer l'impact de la variation du montant octroyé à des étudiants dans le cadre d'un programme de bourses d'études visant à renforcer le taux de scolarisation au niveau secondaire. Une bourse de 60 dollars est accordée aux 25 élèves de chaque école

obtenant les meilleurs résultats à la fin du cycle primaire, et une bourse de 45 dollars est accordée aux 25 suivants. Les élèves obtenant les moins bons résultats ne reçoivent pas de bourse. Dans ce contexte, un modèle de discontinuité de la régression permet de comparer les résultats des élèves non seulement autour du seuil de 45 dollars, mais également autour du seuil de 60 dollars. Filmer et Schady (2009) présentent les résultats d'une évaluation de ce type réalisée au Cambodge à l'issue de laquelle ils concluent que l'impact de la bourse de 60 dollars sur le taux de scolarisation n'est pas plus élevé que celui de la bourse de 45 dollars. Ce résultat est très important d'un point de vue politique, car il suggère qu'il est possible d'augmenter la couverture du programme d'un tiers avec un même budget (par exemple, distribuer 20 000 bourses de 45 dollars au lieu de 15 000 bourses de 60 dollars) tout en assurant l'efficacité du programme.

Évaluation de traitements multiples à l'aide d'études croisées

Outre la comparaison de différentes intensités de traitement, il est également possible de comparer différents types de traitement. En pratique, les décideurs préfèrent généralement pouvoir comparer les avantages relatifs de différentes interventions plutôt que de connaître l'impact d'une seule intervention.

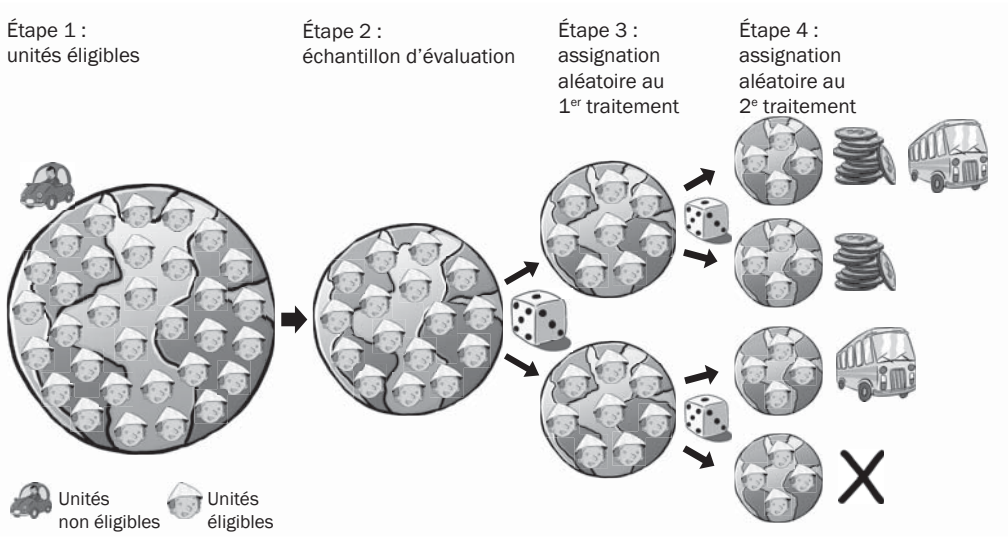
Imaginez que vous voulez évaluer l'impact sur le taux de scolarisation d'un programme comportant deux interventions : des transferts monétaires conditionnels aux familles des étudiants et le transport scolaire gratuit. Vous souhaitez connaître l'impact distinct de chaque intervention et savoir si la combinaison des deux serait plus efficace que la somme des impacts individuels. Le programme est proposé aux participants sous trois formes différentes : les transferts monétaires conditionnels uniquement, le transport scolaire gratuit uniquement ou une combinaison des deux.

L'assignation aléatoire d'un programme comportant deux interventions est comparable à celle utilisée pour les programmes n'en comportant qu'une seule. La principale différence réside dans la nécessité d'effectuer plusieurs tirages au sort au lieu d'un seul dans ce qui s'appelle une *étude croisée*. La figure 9.2 illustre ce processus. Comme indiqué précédemment, la première étape consiste à définir les unités éligibles au programme. La deuxième étape consiste à sélectionner un échantillon d'unités à partir de la population pour constituer l'échantillon d'évaluation. Une fois l'échantillon d'évaluation créé, la troisième étape consiste à répartir de façon aléatoire les unités entre le groupe de traitement et le groupe de comparaison. À l'étape 4, vous effectuez un deuxième tirage au sort pour sélectionner de façon aléatoire les unités du groupe de traitement qui bénéficieront de la première intervention. Enfin, vous effectuez un autre tirage au sort pour sélectionner un sous-groupe au sein du groupe de comparaison qui bénéficiera de la deuxième intervention, le reste du sous-groupe restant entièrement à l'écart des interventions.

Le processus d'assignation aléatoire appliqué aux deux traitements permet d'obtenir quatre groupes, comme l'illustre la figure 9.3.

- Le groupe A bénéficie des deux interventions (transferts monétaires et transport scolaire).

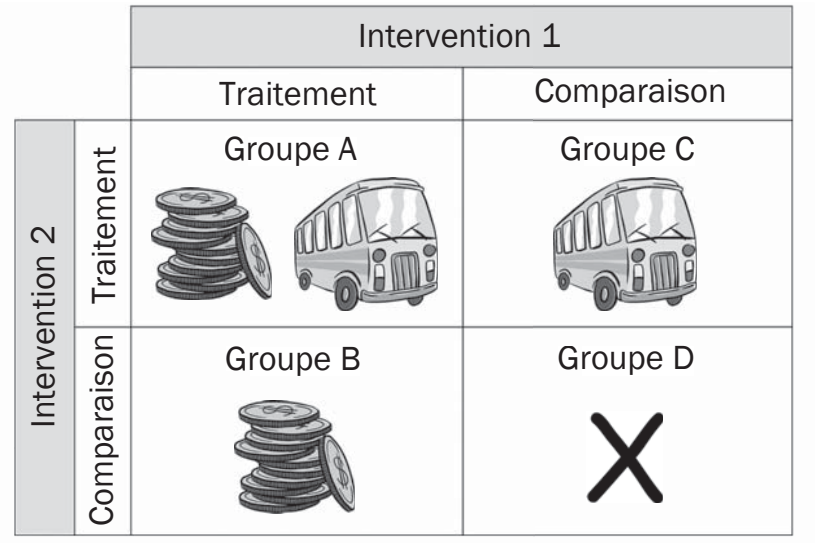
Figure 9.2 Étapes de l'assignation aléatoire pour deux interventions



- Le groupe B bénéficie uniquement de l'intervention 1 (transferts monétaires).
- Le groupe C bénéficie uniquement de l'intervention 2 (transport scolaire).
- Le groupe D ne bénéficie d'aucune des deux interventions et constitue le groupe de comparaison « pur ».

Si elle est correctement effectuée, l'assignation aléatoire permet de créer quatre groupes similaires. Vous pouvez alors estimer l'impact de la première intervention en comparant le résultat du groupe B à celui du groupe D, le groupe de comparaison « pur ». Vous pouvez également estimer l'impact de la deuxième intervention en comparant le résultat du groupe C à celui du groupe de comparaison non exposé. Ce processus permet également de comparer l'impact supplémentaire lié à l'assignation de la deuxième intervention sur les unités bénéficiant déjà de la première. En comparant les résultats du groupe A et du groupe B, vous obtenez l'impact de la deuxième intervention sur les unités qui bénéficient déjà de la première ; et en comparant les résultats du groupe A et du groupe C, on obtient l'impact de la première intervention sur les unités qui bénéficient de la deuxième.

Figure 9.3 Groupes de traitement et groupe de comparaison pour un programme à deux interventions



L'explication ci-dessus se réfère au cas de l'assignation aléatoire pour décrire comment élaborer une évaluation d'impact pour un programme comportant deux interventions. Lorsqu'un programme comporte plus de deux interventions, il est possible d'augmenter le nombre de tirages au sort et continuer à subdiviser la population pour créer des groupes soumis à différentes combinaisons d'interventions. Il est également envisageable de réaliser des évaluations combinant plusieurs traitements et plusieurs niveaux de traitement. Même si le nombre de groupes augmente, la théorie reste la même.

Toutefois, l'évaluation de plusieurs interventions peut présenter des difficultés pratiques à la fois au stade de l'évaluation que de la mise en œuvre du programme. En effet, le programme est plus complexe et le nombre de branches du traitement augmente exponentiellement. Pour l'évaluation d'une intervention, seuls deux groupes sont nécessaires : le groupe de traitement et le groupe de comparaison. Pour l'évaluation de deux interventions, quatre groupes sont nécessaires : trois groupes de traitement et un groupe de comparaison. Pour évaluer trois interventions en tenant compte de toutes les combinaisons possibles entre ces interventions, il faut $2 \times 2 \times 2 = 8$ groupes. En résumé, pour qu'une évaluation couvre toutes les combinaisons possibles entre n interventions, il faut 2^n groupes. Par ailleurs, pour être en

Encadré 9.1 : Comparer des alternatives de programmes de prévention du VIH/sida au Kenya

Duflo et al. (2006) évaluent l'impact de plusieurs programmes de prévention du VIH/sida dans deux zones rurales à l'Ouest du Kenya à l'aide d'une étude croisée. L'étude est basée sur un échantillon de 328 écoles réparties en six groupes, comme l'illustre le tableau ci-dessous qui résume la mise en œuvre du programme. Chaque groupe bénéficie d'une combinaison différente de trois traitements assignés de façon aléatoire. Les traitements comprennent un programme de formation des enseignants visant à renforcer leurs capacités à enseigner le programme national d'éducation sur le VIH/sida, la promotion de l'organisation de débats sur le rôle des préservatifs dans les écoles et l'organisation de concours de rédaction sur le thème de la prévention, ainsi que la réduction des frais d'éducation grâce à la distribution gratuite d'uniformes scolaires (voir tableau).

Résumé de la mise en œuvre du programme

Groupe	Nombre d'écoles	Programme national	Formation des enseignants	Débat sur les préservatifs et rédaction d'essais (printemps 2005)	Baisse des frais d'éducation (printemps 2003 et automne 2004)
1	88	Oui			
2	41	Oui	Oui		
3	42	Oui	Oui	Oui	
4	83				Oui
5	40	Oui	Oui		Oui
6	40	Oui	Oui	Oui	Oui

Les chercheurs concluent qu'au bout de deux ans, le programme de formation des enseignants n'a qu'un impact limité sur les connaissances des élèves, les activités sexuelles rapportées, l'utilisation du préservatif ou les grossesses chez les adolescentes, bien qu'il ait amélioré l'enseignement du programme national. Les débats et les concours de rédaction renforcent les connaissances et l'utilisation des préservatifs sans augmenter les activités sexuelles rapportées. Enfin, la réduction des frais d'éducation de par la distribution d'uniformes scolaires permet de réduire les taux d'abandon et les grossesses chez les adolescentes. Les chercheurs concluent que la distribution d'uniformes scolaires a un impact plus marqué sur la réduction des grossesses chez les adolescentes que la formation des enseignants au programme national sur le VIH/sida.

Source : Duflo et al. 2006.

Encadré 9.2 : Comparer différents programmes de suivi de la corruption en Indonésie

En Indonésie, Olken (2007) utilise une étude croisée novatrice pour étudier différentes méthodes de contrôle de la corruption, à savoir des audits gouvernementaux et un suivi communautaire sur le terrain. Il applique la méthodologie de l'assignation aléatoire dans plus de 600 villages où des routes allaient être construites dans le cadre d'un projet national d'amélioration des infrastructures.

D'une part, l'un des traitements consiste à prévenir certains villages sélectionnés de manière aléatoire que leur projet de construction allait faire l'objet d'un audit par un agent gouvernemental. D'autre part, pour mesurer la participation de la communauté au contrôle de la corruption, les chercheurs mettent en place deux interventions. Ils organisent des réunions de responsabilisation communautaire et distribuent des fiches de commentaires pouvant être remplis de façon anonyme. Pour mesurer les niveaux de corruption, une équipe indépendante d'ingénieurs et d'arpenteurs prélève des échantillons des nouvelles routes, estimant le coût des matériaux utilisés puis comparant les résultats obtenus aux budgets déclarés.

Olken conclut que l'augmentation des audits gouvernementaux (la probabilité d'audit passant d'environ 4 % à 100 %) permet de réduire les dépenses manquantes d'environ huit points (d'un point de départ de 24 %). L'intensification de la participation communautaire au contrôle de la corruption exerce un impact sur l'absence des ouvriers, mais pas sur les dépenses manquantes. Les fiches de commentaires ne donnent des résultats probants que lorsqu'elles sont distribuées aux enfants à l'école pour être remis à leurs parents et non quand ils sont distribués par les chefs de village.

Source : Olken, 2007.

mesure de distinguer les différences de résultats entre les différents groupes, chaque groupe doit contenir un nombre suffisant d'unités pour garantir une puissance statistique satisfaisante. Afin de déceler des différences entre les différentes branches de l'intervention, des échantillons plus importants seront nécessaires que pour effectuer de simples comparaisons d'un groupe de traitement et d'un groupe de comparaison. Si les deux branches du traitement entraînent des changements de résultats, des échantillons plus importants devront être constitués pour détecter d'éventuelles différences (souvent plus petites) entre les deux groupes.

Finalement, les études croisées peuvent également être mises en place dans le cadre d'évaluations combinant plusieurs méthodes (encadrés 9.1 et 9.2). Les règles opérationnelles qui régissent l'assignation de chaque traitement déterminent la combinaison des méthodes à utiliser. Par exemple, le premier traitement peut être attribué sur la base d'un seuil d'éligibilité tandis que le deuxième est assigné de manière aléatoire. Dans ce cas, il est possible de réaliser un modèle de discontinuité de la régression pour la première intervention et suivre une méthode d'assignation aléatoire pour la seconde.

Note

1. Voir Banerjee et Duflo (2009) pour une explication plus détaillée.

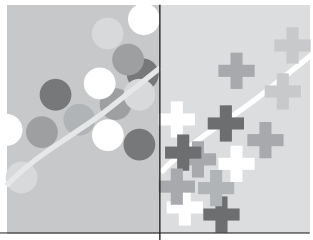
Références

Banerjee, Abhijit et Esther Duflo. 2009. « The Experimental Approach to Development Economics. » NBER Working Paper 14467, National Bureau of Economic Research, Cambridge, MA.

Duflo, Esther, Pascaline Dupas, Michael Kremer et Sameul Sinei. 2006. « Education and HIV/AIDS Prevention: Evidence from a Randomized Evaluation in Western Kenya. » Document de travail consacré à la recherche sur les politiques 402, Banque mondiale, Washington, DC.

Filmer, Deon et Norbert Schady. 2009. « School Enrollment, Selection and Test Scores. » Document de travail consacré à la recherche sur les politiques 4998, Banque mondiale, Washington, DC.

Olken, Benjamin. 2007. « Monitoring Corruption: Evidence from a Field Experiment in Indonesia. » *Journal of Political Economy* 115 (2) : 200–49.



Partie 3

COMMENT METTRE EN ŒUVRE UNE ÉVALUATION D'IMPACT

Dans la première partie de l'ouvrage, nous avons exposé pourquoi effectuer des évaluations d'impact et expliqué quand elles sont opportunes. Les évaluations sont conçues pour répondre à des questions de politique bien définies, par exemple, dans le cadre de négociations budgétaires ou pour prendre des décisions sur l'extension d'un programme alimentaire, l'augmentation du montant de bourses pour les étudiants ou la mise en œuvre d'une réforme hospitalière. Les objectifs de l'évaluation et les questions qui l'orientent doivent découler directement de ces questions politiques. Après avoir clairement défini le programme à évaluer et les questions de politique sur lesquelles l'évaluation doit porter, il est utile d'élaborer une théorie du changement, telle qu'une chaîne de résultats du programme, et de choisir des indicateurs en conséquence. Dans la deuxième partie de cet ouvrage, nous avons décrit une série de méthodes, illustrées par des exemples, qui permettent d'évaluer l'impact d'un programme ; nous avons présenté les avantages et les inconvénients de chacune d'elles.

La troisième partie porte sur les étapes opérationnelles qui jalonnent la gestion ou la commande d'une évaluation d'impact. Ces étapes constituent les éléments clefs de la réalisation d'une évaluation d'impact dans le but de répondre aux questions de politique formulées et d'estimer l'impact causal du programme. Les étapes opérationnelles d'une évaluation d'impact peuvent être regroupées en quatre phases principales : *conception de l'évaluation, choix d'un échantillon, collecte des données* et *production et diffusion des résultats*. La figure ci-dessous illustre ces phases, détaillées dans les chapitres 10 à 13.

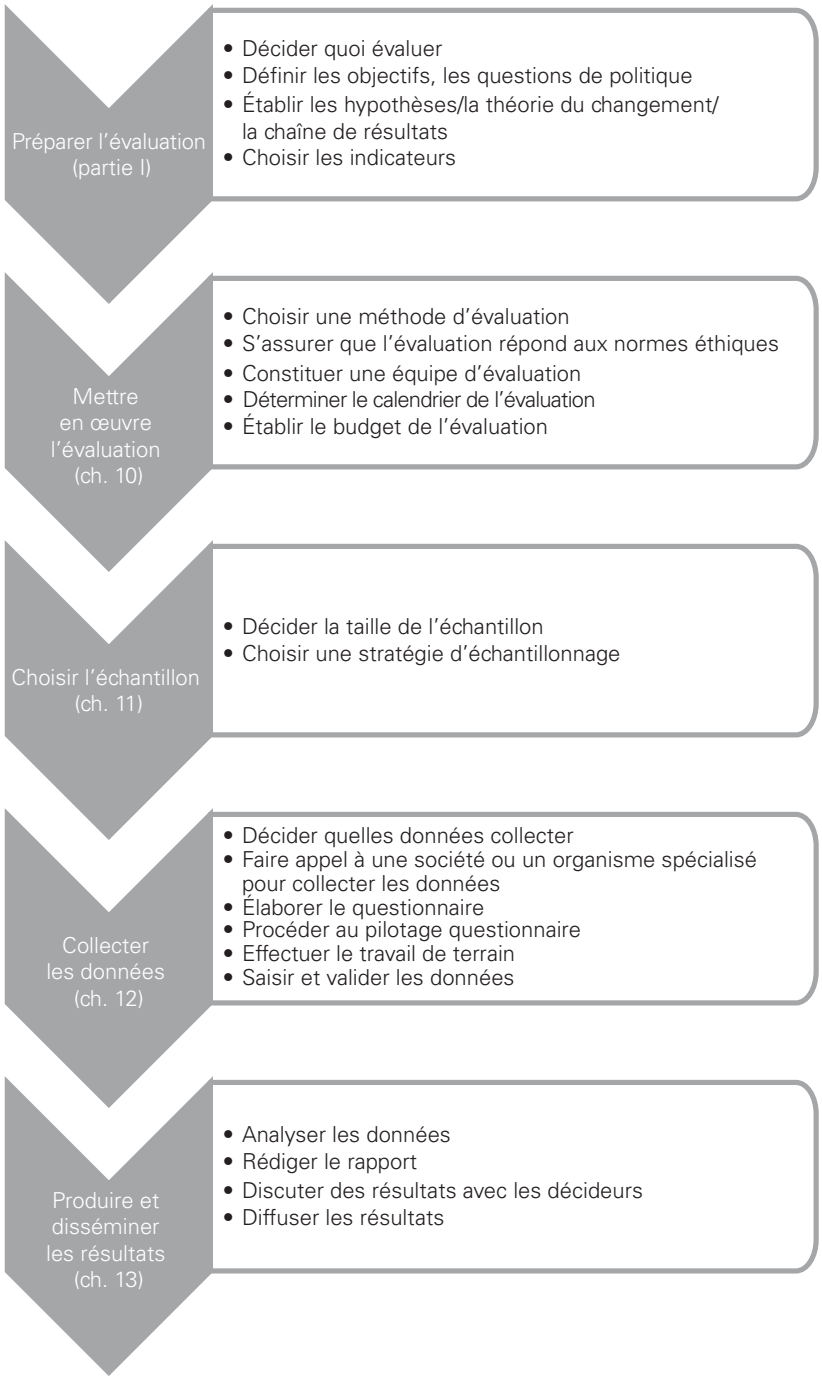
Le chapitre 10 porte sur les principales composantes de la mise en œuvre d'une évaluation. Elle commence par le choix d'une méthode d'évaluation en fonction du plan d'implémentation du programme. Avant de pouvoir mettre l'évaluation en œuvre, vous vous assurez qu'elle répond à des normes d'éthique. Vous constituez ensuite une équipe chargée de l'évaluation, vous établissez un budget et définissez un mode de financement.

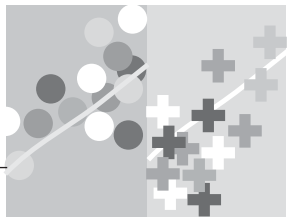
Le chapitre 11 passe en revue comment constituer des échantillons pour collecter des données et combien d'unités y inclure.

Au chapitre 12, nous abordons les différentes étapes de la collecte de données. En gardant à l'esprit les questions de politique auxquelles vous voulez répondre et la conception de votre évaluation, vous devez déterminer si les données existantes sont suffisantes et décider si de nouvelles données doivent être collectées. Vous commanditez la rédaction d'un questionnaire pertinent qui vous permettra de mesurer les indicateurs choisis. Vous choisissez ensuite une entreprise ou un organisme spécialisé en collecte de données. Celui-ci recrutera et formera du personnel de terrain et procédera au pilotage du questionnaire. Après avoir effectué les ajustements nécessaires, l'entreprise ou l'organisme pourra entamer le travail sur le terrain. Enfin, les données collectées sont saisies ou traitées et validées avant d'être analysées.

Le chapitre 13 porte sur les étapes finales de l'évaluation. Il décrit les produits générés par l'évaluation et le contenu des rapports d'évaluation, et énonce quelques lignes directrices sur la manière de diffuser les résultats auprès des décideurs et des différentes parties prenantes.

Figure P3.1 Feuille de route pour la mise en œuvre d'une évaluation d'impact





CHAPITRE 10

Mettre en œuvre une évaluation d'impact

Dans la deuxième partie de ce manuel, nous avons évoqué diverses méthodes permettant de générer des groupes de comparaison valides. L'estimation de l'impact causal d'un programme est fondée sur ces groupes de comparaison. Nous allons maintenant examiner les aspects pratiques relatifs au choix de la méthode la plus adéquate pour un programme donné. Comme nous le verrons, les règles opérationnelles du programme déterminent la provenance des groupes de comparaison et, partant, la méthode d'évaluation la plus appropriée compte tenu du contexte.

Choisir une méthode d'évaluation

La validité de l'estimation de l'impact causal d'un programme dépend essentiellement de l'existence d'un groupe de comparaison de qualité. Dans la deuxième partie de ce manuel, nous avons évoqué plusieurs groupes de comparaison valides, en particulier ceux générés par l'assignation aléatoire, la promotion aléatoire, le modèle de discontinuité de la régression, la méthode de la double différence et l'appariement. Dans le présent chapitre, nous considérons comment choisir l'une de ces méthodes en fonction du contexte. Le principe fondamental est que les règles opérationnelles du programme permettent de déterminer la méthode la mieux adaptée : ce sont donc ces règles qui doivent conduire à la méthode d'évaluation, et non l'inverse. La mise en place d'une évaluation ne doit en aucun cas requérir des changements radicaux des éléments clés de l'intervention dans le seul but d'utiliser une méthode d'évaluation donnée.

Concept clé :
Les règles opérationnelles du programme déterminent le choix de la méthode d'évaluation (et non l'inverse).

L'assignation aléatoire est souvent la méthode préférée des évaluateurs. Appliquée dans les règles de l'art, elle assure la comparabilité tant des caractéristiques observables que des caractéristiques non observables du groupe de traitement et du groupe de comparaison, tout en présentant un risque de biais limité. Puisque l'assignation aléatoire fournit une estimation de l'effet de traitement moyen sur une population donnée de manière largement intuitive et exige peu de connaissances en économétrie, elle facilite grandement la communication des résultats aux décideurs. Il n'est toutefois pas toujours possible d'utiliser des méthodes d'assignation aléatoire, notamment lorsqu'elles sont incompatibles avec les règles opérationnelles des programmes.

Les règles opérationnelles les plus importantes pour la conception d'une évaluation sont celles qui permettent d'identifier les unités éligibles à un programme et la manière dont s'effectue la sélection en vue de la participation au programme. Les groupes de comparaison sont constitués à partir de la population éligible qui ne peut pas être intégrée au programme à un moment donné (par exemple, si la demande est supérieure à l'offre) ou à partir de la population qui présente des caractéristiques proches de celles nécessaires pour participer à un programme, mais qui en est exclue en raison des règles de ciblage ou d'éligibilité du programme. Il est difficile de trouver des groupes de comparaison valides si les règles régissant l'éligibilité et la sélection ne sont pas équitables, transparentes et que les gestionnaires des programmes en sont tenus responsables.

Principes des règles de ciblage

Il est pratiquement toujours possible de déterminer un groupe de comparaison valide si les règles opérationnelles de sélection des bénéficiaires sont équitables, transparentes et que les gestionnaires des programmes en sont tenus responsables :

- Les règles *équitables* définissent un classement ou un ordre d'éligibilité selon un indicateur reconnu des besoins, ou offrent le programme à tous, ou du moins donnent à tous une chance égale d'en bénéficier.
- Le critère de *transparence* implique que les règles soient publiques de manière à ce que la société civile puisse les reconnaitre et vérifier qu'elles sont bien respectées. Pour être transparentes, les règles doivent être quantitatives et faciles à observer par les parties externes.
- Les gestionnaires des programmes sont *tenus responsables* des règles de sélection des bénéficiaires quand ils doivent en rendre compte et quand elles constituent la base sur laquelle leur performance est mesurée et leur rétribution établie.

Concept clé :
Il est pratiquement toujours possible de déterminer un groupe de comparaison valide si les règles opérationnelles de sélection des bénéficiaires sont équitables, transparentes et que les gestionnaires des programmes en sont tenus responsables.

Comme nous le verrons ultérieurement, le critère d'équité implique souvent l'utilisation soit de l'assignation aléatoire, soit du modèle de discontinuité de la régression. La transparence et la responsabilisation des gestionnaires permettent de s'assurer que les critères de ciblage sont quantitativement vérifiables et mis en œuvre comme prévu. Si les règles opérationnelles ne respectent pas ces trois principes de bonne gouvernance, tant la conception du programme que la mise en œuvre de l'évaluation deviennent problématiques.

Les règles opérationnelles d'éligibilité répondent aux critères de transparence et de responsabilisation quand elles sont quantifiables, publiques, et qu'elles peuvent faire l'objet d'une vérification externe. Ces principes de bonne gouvernance augmentent la probabilité qu'un programme bénéficie réellement à la population cible et constituent la pierre angulaire d'une bonne évaluation. Si les règles ne sont ni quantifiables ni vérifiables, l'équipe chargée de l'évaluation aura du mal à vérifier si l'assignation au groupe de traitement ou au groupe de comparaison a été effectuée comme prévu ou même à comprendre comment cette assignation a eu lieu. Si les évaluateurs sont dans l'incapacité d'observer le processus d'assignation, ils ne pourront pas analyser correctement les données pour déterminer l'impact du programme. La compréhension des règles d'assignation du programme est absolument essentielle pour déterminer la méthode d'évaluation d'impact la plus adéquate.

Règles opérationnelles de ciblage

Les règles opérationnelles régissent les bénéfices offerts par le programme, leur financement et leur distribution, ainsi que le mode de sélection des bénéficiaires. Les règles de financement du programme et d'intégration des bénéficiaires sont fondamentales à la définition des groupes de comparaison valides. Les règles d'intégration des bénéficiaires recouvrent l'éligibilité, l'allocation de ressources limitées et le calendrier d'intégration des bénéficiaires. Plus précisément, les règles clés pour choisir les groupes de comparaison relèvent de trois questions opérationnelles fondamentales, se rapportant elles-mêmes au financement, au ciblage et au calendrier :

1. *Financement : le programme dispose-t-il de suffisamment de ressources pour couvrir l'ensemble de la population éligible ?* Les organismes publics et les organisations non gouvernementales ne disposent pas toujours des fonds nécessaires pour offrir les services du programme à toutes les personnes éligibles qui souhaitent y participer. Dans de tels cas, les autorités doivent décider qui, parmi la population éligible, intégrera le programme et qui en sera exclu. Les programmes sont parfois limités à une région donnée, aux zones rurales ou à de petites communautés même s'il existe des personnes éligibles dans d'autres régions ou dans des communautés plus grandes.

2. *Ciblage : qui est éligible au programme ? Un seuil d'éligibilité a-t-il été fixé ou le programme est-il ouvert à tout le monde ?* L'instruction publique et les services de santé primaires sont généralement offerts à tous. Toutefois, de nombreux programmes s'appuient sur des règles de ciblage reposant sur un classement continu des bénéficiaires potentiels et la définition d'un seuil d'éligibilité. Par exemple, les programmes de retraites fixent un âge à partir duquel les personnes deviennent éligibles. Les programmes de transferts monétaires classent souvent les ménages selon leur niveau de pauvreté, et seuls les ménages au-dessous d'un certain seuil sont considérés comme éligibles.

3. *Calendrier : comment les bénéficiaires potentiels sont-ils intégrés au programme — tous ensemble en une seule fois ou par phases à travers le temps ?* Dans de nombreux cas, les contraintes administratives et de ressources empêchent les autorités de servir immédiatement tous les membres du groupe cible. Si le programme doit être mis en œuvre par phases, il faut alors désigner qui seront les premiers bénéficiaires et qui seront les suivants. Une approche commune consiste à élargir le programme par région, en intégrant tout d'abord les populations éligibles d'un village ou d'une région donnés, puis progressivement les autres.

Identification et classement des bénéficiaires par ordre de priorité

Les trois questions précédentes se réfèrent à l'aspect opérationnel fondamental du mode de sélection des bénéficiaires. Comme nous le verrons plus tard, cet aspect est essentiel pour définir des groupes de comparaison valides. Les groupes de comparaison proviennent soit de la population non éligible soit, plus fréquemment, de la population éligible non encore intégrée au programme. La manière dont l'ordre d'intégration est établi dépend en partie des objectifs du programme. S'agit-il d'un programme de retraites pour les personnes âgées, d'un programme de réduction de la pauvreté, ou d'un programme de vaccination ouvert à l'ensemble de la population ?

Pour déterminer l'ordre de priorité des bénéficiaires, la définition d'un indicateur à la fois quantifiable et vérifiable est nécessaire. Lorsqu'un indicateur est établi, son application dépend de la capacité des autorités à mesurer les besoins et à les classer par ordre de priorité. Si les autorités peuvent précisément classer les bénéficiaires potentiels en fonction de leurs besoins relatifs, un déploiement du programme selon l'ordre de priorité déterminé par ces besoins répond à des raisons éthiques. Toutefois, pour établir un tel classement en fonction des besoins, il faut non seulement disposer d'un indicateur quantifiable, mais avoir les capacités et les ressources suffisantes pour procéder à des mesures individuelles de cet indicateur.

Dans certains cas, l'éligibilité à un programme est déterminée par un indicateur continu, pour lequel la collecte des données est facile et bon marché, par exemple l'âge d'admissibilité dans le cadre de prestations de retraite. Par exemple, l'âge de 70 ans constitue un seuil d'éligibilité à la retraite simple à mesurer et à appliquer. Pourtant, ce type d'indicateur ne permet le plus souvent pas de classer les besoins relatifs au sein de la population éligible. Par exemple, une personne de 69 ans n'a pas forcément moins besoin de prestations de retraite qu'une personne de 70 ans ; de la même manière, une personne de 75 ans n'a pas forcément plus besoin d'une retraite qu'une personne de 72 ans. Avec un indicateur comme l'âge de la retraite, il est possible d'identifier la population éligible, mais il n'est pas nécessairement possible d'établir un classement des besoins relatifs au sein de cette population.

D'autres programmes établissent des critères d'éligibilité qui pourraient aussi à priori permettre de déterminer l'éligibilité et d'établir un classement des besoins relatifs. Par exemple, de nombreux projets visent les populations pauvres, mais les indicateurs de pauvreté fiables qui permettent de classer les ménages sont souvent difficiles à mesurer, et la collecte des données nécessaires est souvent onéreuse.

La collecte de données sur le revenu ou la consommation de l'ensemble des bénéficiaires potentiels dans le but de les classer par niveau de pauvreté constitue un processus complexe et onéreux. Pour cette raison, nombreux sont les programmes qui utilisent une approche indirecte comme un test de type « proxy mean » pour estimer le niveau de pauvreté. Ces approches fournissent des mesures approximatives du niveau de pauvreté des bénéficiaires potentiels en se fondant sur leur possession d'actifs ou leurs caractéristiques sociodémographiques (Grosh et al. 2008). Cependant, ces mesures peuvent contenir des erreurs, coûtent cher et ne permettent pas toujours d'établir un classement précis des ménages selon leurs besoins ou leur statut socioéconomique, surtout dans la partie inférieure de la distribution du revenu. Les tests de type « proxy mean » peuvent contribuer à déterminer de manière relativement fiable si un ménage donné se situe au-dessus ou au-dessous d'un seuil donné, mais se révèlent moins efficaces lorsqu'il s'agit d'estimer la distance par rapport à ce seuil. Ces approches permettent d'identifier les populations pauvres éligibles, mais pas forcément d'établir un classement de ces populations en fonction de leurs besoins relatifs.

Pour contourner les problèmes de coûts et la complexité associés au classement des individus ou des ménages selon leurs besoins relatifs, le ciblage des programmes s'effectue souvent à un niveau supérieur, par exemple au niveau des communautés. L'hypothèse sous-jacente à cette approche est que les ménages qui composent les communautés sont globalement homogènes et que la grande majorité de la population est potentiellement éligible. Il serait dès lors injustifié de subir des couts élevés dans le seul but d'identifier un nombre limité d'individus inéligibles. Dans ce cas, tous les membres de la communauté sont considérés comme éligibles au programme. Cette stratégie est souvent efficace pour de petites communautés rurales, mais elle l'est moins pour les programmes réalisés en zones urbaines, où les populations sont plus hétérogènes. Le ciblage à un niveau d'agrégation élevé présente des avantages opérationnels indéniables, mais ne permet pas toujours d'éviter le classement des bénéficiaires sur la base d'un indicateur objectif et quantifiable des besoins.

Si l'agence qui assure le financement du programme décide de ne pas établir de classement des besoins, car elle estime le risque d'erreur ou les coûts trop élevés, elle doit recourir à d'autres critères pour définir comment articuler la séquence des différentes phases du programme. L'équité est un critère compatible avec les principes de bonne gouvernance. Une règle équitable peut consister à donner à toutes les personnes éligibles la même chance d'être intégrées dans la première phase du programme et d'assigner, de manière aléatoire, les bénéficiaires potentiels à l'une des phases suivantes du programme. Cette règle d'allocation est non seulement juste et équitable, mais elle permet de garantir la validité interne et externe de l'évaluation.

Passer des règles opérationnelles aux groupes de comparaison

Dans le tableau 10.1, nous présentons les groupes de comparaison possibles en fonction des règles opérationnelles des programmes et des trois questions opérationnelles fondamentales relatives au financement, au ciblage et au calendrier que nous avons évoquées ci-dessus. Le tableau comprend deux colonnes principales : la première correspond aux cas où le programme n'est pas doté des ressources suffisantes pour couvrir l'ensemble des bénéficiaires potentiels, et la seconde aux cas où ces ressources

Tableau 10.1 Relations entre les règles opérationnelles d'un programme et les méthodes d'évaluation d'impact

	FINANCEMENT →	**Demande supérieure à l'offre (ressources limitées)**		**Pas de demande excédentaire (ressources suffisantes)**	
	RÈGLES DE CIBLAGE →	**Ciblage selon classement continu et seuil d'éligibilité (1)**	**Pas de ciblage selon classement continu et seuil d'éligibilité (2)**	**Pas de ciblage selon classement continu et seuil d'éligibilité (3)**	**Pas de ciblage selon classement continu et seuil d'éligibilité (4)**
CALENDRIER	**Mise en œuvre par phases (A)**	**CELLULE A1** (3.1) Assignation aléatoire (4) MDR	**CELLULE A2** (3.1) Assignation aléatoire (3.2) Promotion aléatoire (5) DD avec (6) Appariement	**CELLULE A3** (3.1) Assignation aléatoire par phases (4) MDR	**CELLULE A4** (3.1) Assignation aléatoire par phases (3.2) Promotion aléatoire pour participation à phase initiale (5) DD avec (6) Appariement
	Mise en œuvre immédiate (B)	**CELLULE B1** (3.1) Assignation aléatoire (4) MDR	**CELLULE B2** (3.1) Assignation aléatoire (3.2) Promotion aléatoire (5) DD avec (6) Appariement	**CELLULE B3** (4) MDR	**CELLULE B4** En absence de participation universelle : (3.2) Promotion aléatoire (5) DD avec (6) Appariement

Remarque : les chiffres entre parenthèses renvoient au chapitre du manuel où la méthode est présentée. MDR = modèle de discontinuité de la régression ; DD = double différence

sont suffisantes (*financement*). Chacune de ces deux colonnes est à son tour subdivisée en deux autres colonnes selon que le programme est ciblé ou ouvert à tous (*règles de ciblage*). Les lignes sont divisées en fonction des impératifs temporels (*calendrier*), selon que les bénéficiaires du programme sont intégrés immédiatement au programme ou par phases. Chaque cellule du tableau indique les méthodes possibles pour former un groupe de comparaison valide. Chaque cellule est associée à une lettre indiquant sa place dans les lignes du tableau (ligne A ou B) et à un chiffre représentant les colonnes (de 1 à 4). Par exemple la cellule A1 se réfère à la première ligne (A) et à la première colonne (1). Dans la cellule A1 figurent les méthodes d'évaluation les plus adaptées aux programmes ciblés, dotés de ressources limitées et mis en œuvre par phases.

Pour la plupart des programmes, une mise en œuvre par phases est nécessaire du fait de contraintes financières, logistiques ou administratives. Cette catégorie de programmes se retrouve dans la première ligne du tableau (cellules A1, A2, A3 et A4). Dans ces cas-là, la règle opérationnelle la plus équitable, la plus transparente et qui permet de tenir les gestionnaires des programmes responsables consiste à donner à chacun une chance égale d'intégrer le programme dans chacune des phases, autrement dit de procéder par assignation aléatoire aux diverses phases du programme.

Lorsque les ressources sont limitées, c'est-à-dire dans les cas où les ressources sont insuffisantes pour couvrir l'ensemble de la population (cellules A1 et A2, ainsi que B1 et B2), la demande peut rapidement dépasser l'offre. Un tirage au sort est alors un bon moyen de choisir les bénéficiaires parmi une population ayant les mêmes besoins relatifs. Ainsi, chacun a une chance égale d'intégrer le programme. Le tirage au sort est une règle opérationnelle d'allocation des services d'un programme qui est équitable, transparente et qui permet de tenir les gestionnaires des programmes responsables.

Les cellules A1 et A3 comprennent une autre catégorie de programmes, à savoir ceux qui doivent être mis en œuvre par phases et où un classement des bénéficiaires selon les besoins est possible. Si les bénéficiaires potentiels sont classés selon des critères quantitatifs et qu'un seuil d'éligibilité peut être fixé, un modèle de discontinuité de la régression peut être adopté.

Les cellules de la dernière ligne du tableau regroupent une autre grande catégorie, les programmes pour lesquels les capacités administratives sont suffisantes pour permettre une mise en œuvre immédiate. Lorsque les ressources sont limitées et qu'il n'est pas possible d'établir un classement des bénéficiaires (cellule B2), l'évaluation peut avoir recours à une assignation aléatoire quand la demande est supérieure à l'offre. Si les ressources sont suffisantes pour couvrir l'ensemble de la demande et qu'il n'y pas de critères de ciblage (cellule B4), la promotion aléatoire est alors la seule possibilité pour autant que la participation au programme ne soit pas universelle. S'il est possible d'établir une priorité parmi les bénéficiaires potentiels et que le programme est ciblé, le modèle de discontinuité de la régression peut de nouveau faire l'affaire.

Détermination de l'échelle minimum de l'intervention

Les règles opérationnelles déterminent également l'échelle minimum d'intervention, c'est-à-dire le niveau auquel le programme est mis en œuvre. Par exemple, si un programme de santé est exécuté à l'échelle régionale, tous les villages de la région en bénéficieront (en groupe) ou en seront exclus. Certains programmes peuvent être efficacement mis en œuvre au niveau des individus, des ménages ou des institutions tandis que d'autres doivent être implémentés au niveau d'une communauté ou d'une région administrative. L'exécution d'une intervention à un niveau élevé (par exemple au niveau d'une province ou d'un État) peut se révéler problématique pour l'évaluation pour trois raisons principales :

1. La taille de l'échantillon d'évaluation et le coût de l'évaluation augmentent avec l'échelle d'intervention.

2. Plus l'échelle d'intervention est élevée, plus il est difficile de disposer d'un nombre suffisant d'unités à inclure dans l'évaluation.

3. La validité interne de l'évaluation peut être plus à risque avec des unités d'intervention à grande échelle.

Premièrement, les évaluations portant sur des niveaux d'intervention élevés comme des communautés ou des régions administratives exigent des échantillons de taille plus importante et sont plus coûteuses que les évaluations concernant des unités d'un niveau moindre comme les personnes ou les ménages[1]. Le niveau d'intervention est important, car il définit l'unité à laquelle le traitement sera appliqué ainsi que les groupes de comparaison formés, ce qui détermine aussi la taille de l'échantillon d'évaluation et donc son coût. Pour les interventions à un niveau élevé, un échantillon plus important est nécessaire pour pouvoir déterminer l'impact réel du programme. L'intuition sous-jacente à cette affirmation sera examinée au chapitre 11, lequel porte sur les calculs de puissance et la manière de définir la taille de l'échantillon d'évaluation.

Un point légèrement différent est que la taille de l'échantillon nécessaire pour que l'assignation aléatoire génère des groupes de traitement et de comparaison équilibrés devient problématique à des niveaux élevés d'agrégation. Intuitivement, si le niveau d'agrégation est la province et que le pays ne compte que six provinces, l'assignation aléatoire a peu de chances de conduire à des groupes de traitement et de comparaison équilibrés. Supposons que nous affectons trois provinces au groupe de traitement et les trois autres au groupe de comparaison ; il est très peu probable que les provinces du groupe de traitement soient similaires à celles du groupe de comparaison même si le nombre de ménages dans chaque province est important. Pour équilibrer les groupes de comparaison et de traitement, l'élément clé est le nombre d'unités affectées à chacun des deux groupes (dans ce cas le nombre de provinces) et non pas le nombre d'individus ou de ménages dans l'échantillon.

Le troisième problème lorsque l'intervention est mise en œuvre à un niveau élevé est que les changements différentiels dans le temps ont plus de risques d'affecter la validité interne de la sélection aléatoire même si les caractéristiques des groupes sont initialement équilibrées. Revenons à notre exemple des provinces comme niveau d'intervention dans le cadre du programme d'assurance maladie. Certaines provinces sont assignées de manière aléatoire au groupe de traitement et d'autres au groupe de comparaison. Supposons que nous avons de la chance et que les deux groupes sont équilibrés au départ, c'est-à-dire que les ménages du groupe de traitement et ceux du groupe de comparaison affichent initialement des dépenses de santé directes moyennes équivalentes. Après la collecte des données de référence, certaines provinces peuvent décider de lancer d'autres programmes de santé comme des programmes de vaccination ou encore des projets d'approvisionnement en eau et d'assainissement qui permettent d'améliorer la santé de la population et, de ce fait, de réduire les dépenses de santé directes des ménages. Si les groupes de comparaison et de traitement ne bénéficient pas tous des mêmes politiques, l'impact de notre programme d'assurance maladie sur les dépenses de santé directes des ménages se confondra avec l'impact des autres politiques de santé mises en œuvre par certaines

provinces. De même, certaines provinces peuvent enregistrer une croissance économique supérieure à d'autres. Or, les dépenses de santé ont de fortes chances d'augmenter plus rapidement dans les provinces où la croissance est plus importante. Là aussi, si la croissance économique diffère dans les groupes de comparaison et de traitement, l'impact du programme d'assurance maladie sur les dépenses de santé directes risque d'être difficile à isoler de l'impact de la croissance économique sur les économies locales. En général, il est difficile de tenir compte de ces changements lorsqu'ils ont lieu à des niveaux d'intervention élevés. L'assignation aléatoire à des niveaux d'intervention moins élevés permet de mieux maîtriser ces éléments menaçant la cohérence interne de l'évaluation.

Pour éviter les problèmes liés à la mise en œuvre d'une intervention à un niveau géographique ou administratif élevé, les responsables de programme doivent déterminer le niveau minimum auquel le programme peut être mis en œuvre. Cette échelle minimum d'intervention est fonction de plusieurs facteurs :

- Les économies d'échelle et la complexité administrative de la mise en œuvre du programme
- Les capacités administratives de distribuer le programme au niveau des individus ou des ménages
- Les craintes d'éventuels conflits civils
- Les craintes de contamination du groupe de comparaison.

L'échelle minimum d'intervention dépend généralement des économies d'échelle et de la complexité administrative associées à la mise en œuvre du programme. Par exemple, un programme d'assurance maladie peut nécessiter un bureau pour recevoir les demandes des bénéficiaires et régler les fournisseurs. Les coûts fixes de fonctionnement de ce bureau doivent être répartis sur un grand nombre de bénéficiaires ; il peut donc être moins rentable d'exécuter le programme à un niveau individuel qu'au niveau communautaire. Toutefois, lorsqu'il s'agit d'interventions nouvelles non encore éprouvées, il peut être plus judicieux d'accepter les inefficiences à court terme et de mettre en œuvre le programme par district administratif de manière à garantir la crédibilité de l'évaluation et à réduire les coûts de collecte des données.

Les gouvernements argumentent parfois que pour les programmes administrés localement, comme les programmes d'assurance maladie, les capacités administratives sont insuffisantes pour envisager une mise en œuvre au niveau individuel. Ils estiment en effet qu'il serait fastidieux de mettre en place des systèmes pour offrir différents services à différents bénéficiaires à l'intérieur d'unités administratives locales, et qu'il n'est donc pas possible d'effectuer une assignation au groupe de traitement et au groupe de comparaison. Ce problème constitue une sérieuse entrave à la conception de l'évaluation et, de ce fait, à la réussite de l'étude.

Parfois, les autorités préfèrent aussi exécuter les programmes à un niveau d'agrégation plus élevé (par exemple au niveau de la communauté) pour éviter d'éventuels conflits si les membres du groupe de comparaison voient leurs voisins du groupe de traitement bénéficier du programme avant eux. Dans les faits, il existe peu d'élé-

ments pour appuyer ces craintes. De nombreux programmes sont mis en œuvre avec succès au niveau des individus ou des ménages au sein de communautés sans générer de conflit ; il suffit que l'assignation ait lieu de manière équitable, transparente et que les gestionnaires de programme en soient tenus responsables.

D'autre part, lorsqu'un programme est mis en œuvre à un niveau peu élevé, comme l'individu ou le ménage, une contamination du groupe de comparaison peut compromettre la validité interne de l'évaluation. Supposons par exemple que nous cherchons à évaluer l'effet de l'approvisionnement en eau courante sur la santé des ménages. Si des robinets sont installés chez un ménage et pas chez son voisin, le ménage faisant partie du groupe de traitement peut très bien partager l'eau avec son voisin qui, lui, fait partie du groupe de comparaison ; ce voisin ne constituera alors plus un bon point de comparaison du fait de cet effet de débordement.

Dans les faits, les responsables de programme doivent donc trouver l'échelle minimum d'intervention permettant 1) de disposer d'un échantillon d'évaluation suffisamment important, 2) de maîtriser les risques sur le plan de la validité interne, et 3) de s'adapter au contexte opérationnel. L'encadré 10.1 illustre le choix et les implications de l'échelle minimum d'intervention dans le cas des programmes de transferts monétaires.

Encadré 10.1 : Programmes de transferts monétaires et échelle minimum d'intervention

Dans la majorité des programmes de transferts monétaires conditionnels, l'échelle minimum d'intervention est la communauté, pour des raisons administratives et de conception du programme, mais aussi pour éviter les effets de diffusion et d'éventuels conflits qui pourraient naître au sein d'une communauté si le traitement était attribué à un niveau inférieur.

Par exemple, l'évaluation du programme de transferts monétaires conditionnels Progresa/Oportunidades au Mexique repose sur un déploiement du programme au niveau des communautés rurales avec une assignation aléatoire des communautés par phases au groupe de traitement ou au groupe de comparaison. Tous les ménages éligibles des communautés assignées au groupe de traitement intègrent le programme au printemps 1998 et tous les ménages éligibles des communautés assignées au groupe de comparaison l'intègrent 18 mois plus tard, soit à l'hiver 1999. Les évaluateurs trouvent une corrélation importante au niveau des résultats entre les ménages des communautés. Pour garantir une puissance statistique suffisante à l'évaluation, davantage de ménages doivent être inclus dans l'échantillon d'évaluation que ce qui aurait été nécessaire si le groupe de traitement et le groupe de comparaison avaient été constitués au niveau des ménages. L'impossibilité de mettre en œuvre le programme à l'échelle des ménages nécessite donc un échantillon plus grand et entraîne des coûts d'évaluation plus élevés. Ce type de contraintes se retrouve dans un grand nombre de programmes de développement humain.

Sources : Behrman et Hoddinott 2001 ; Gertler 2004 ; Levy et Rodríguez 2005 ; Schultz 2004 ; Skoufias et McClafferty 2001.

L'évaluation est-elle éthique ?

Les évaluations d'impact soulèvent souvent des questions d'éthique. La première question à se poser est de savoir s'il est éthique d'investir des ressources publiques considérables dans des programmes dont l'efficacité n'est pas garantie. Dans ce contexte, c'est plutôt le défaut d'évaluation qui n'est pas éthique. En effet, les informations sur l'efficacité d'un programme produites par les évaluations d'impact peuvent conduire à une utilisation plus efficace et plus éthique des ressources publiques.

Lorsque la décision est prise de mener une évaluation d'impact, d'autres questions d'ordre éthique doivent être considérées. Elles ont trait tant aux règles d'attribution des bénéfices du programme qu'aux méthodes d'étude de sujets humains.

Le premier principe à respecter en matière d'assignation des bénéfices d'un programme est de ne jamais empêcher ou retarder leur distribution à cause de l'évaluation. Dans ce manuel, nous avons déjà souligné que les évaluations ne doivent en aucun cas dicter la manière dont les bénéfices d'un programme sont assignés, mais qu'elles doivent au contraire être adaptées aux règles opérationnelles du programme. Dans ce cadre, les problèmes éthiques qui peuvent survenir ne seront pas liés à l'évaluation d'impact elle-même, mais directement aux règles d'attribution du programme.

Concept clé :
Il ne faut jamais empêcher ou retarder les bénéfices offerts par un programme à cause de l'évaluation.

L'assignation aléatoire des bénéfices du programme pose souvent des questions éthiques liées au fait que certains bénéficiaires éligibles ne participent pas au programme. Pourtant, la plupart des programmes sont dotés de moyens financiers et administratifs limités rendant impossible la couverture immédiate de l'ensemble des bénéficiaires potentiels. D'un point de vue éthique, tous les sujets qui sont également éligibles à la participation à un programme social donné devraient avoir la même chance de bénéficier dudit programme. L'assignation aléatoire répond à ce principe fondamental. Dans les cas où un programme doit être mis en œuvre par phases, une sélection aléatoire peut être effectuée pour déterminer l'ordre selon lequel les personnes formant la population éligible bénéficieront du programme. Les personnes choisies pour intégrer le programme ultérieurement formeront alors le groupe de comparaison, permettant ainsi non seulement de concevoir une bonne étude d'évaluation, mais aussi d'allouer des ressources rares de manière transparente et équitable.

Dans de nombreux pays et institutions internationales, des commissions ou des comités d'éthique ont été mis en place pour encadrer les recherches portant sur les sujets humains. Ces comités sont chargés d'évaluer, d'approuver et de suivre les recherches en cours. Leur objectif premier est de protéger les droits et de promouvoir le bien-être de tous les sujets participant à ces études. Malgré leur orientation opérationnelle, les évaluations d'impact sont également des travaux de recherche et, en tant que telles, doivent se conformer aux directives s'appliquant aux recherches portant sur des sujets humains.

Aux États-Unis, le Bureau de protection de la recherche humaine (Office for Human Research Protections), rattaché au Département de la santé et des services humains (Department of Health and Human Services), est responsable de la coor-

dination des travaux des comités d'éthique institutionnels mis en place dans toutes les universités et institutions de recherche. Ce bureau publie aussi une compilation de plus d'un millier de lois, réglementations et directives relatives au sujet de la recherche humaine dans 96 pays et établit des liens avec les codes éthiques et les normes réglementaires en vigueur dans les principales organisations internationales et régionales.

Par exemple, toutes les recherches menées aux États-Unis ou financées par des agences fédérales américaines comme l'institut national de la santé (National Institutes of Health) ou l'agence américaine de développement international (USAID) doivent être conformes aux principes éthiques et aux exigences réglementaires de la législation fédérale[2]. La législation américaine sur la protection des sujets de recherche humains se base sur le Rapport Belmont et prévoit :

- une sélection équitable des sujets
- la minimisation des risques pour les sujets
- une exposition au risque raisonnable, proportionnelle aux bénéfices attendus
- l'obtention du consentement éclairé de chaque sujet ou de son représentant légal
- l'adoption de dispositions visant à protéger les données personnelles concernant les sujets et à garantir la confidentialité
- la mise en place de dispositions particulières pour protéger les sujets plus vulnérables comme les enfants, les détenus ou les moins nantis.

Les principes élémentaires de protection des droits et de promotion du bien-être de tous les sujets, initialement édictés pour les essais médicaux, s'appliquent aussi aujourd'hui en recherche sociale. Pour l'évaluation des programmes sociaux, les trois premiers points de la liste ci-dessus renvoient aux questions éthiques liées à l'attribution des bénéfices. Les trois derniers concernent les protocoles selon lesquels les sujets humains sont étudiés dans le cadre de l'évaluation[3].

Au moment de concevoir ou de commissionner une évaluation, il convient de bien vérifier que chaque étape est en conformité avec les lois ou procédures d'examen en vigueur qui régissent la recherche sur les sujets humains, que ce soit dans le pays où l'évaluation est effectuée, ou dans le pays de l'organisme qui finance l'évaluation.

Comment constituer une équipe d'évaluation ?

Une évaluation requiert un partenariat entre des décideurs et des évaluateurs, les deux groupes dépendant les uns des autres pour le succès de l'exercice. Les décideurs doivent fournir l'orientation de l'étude et assurer la pertinence de l'évaluation

en déterminant si l'évaluation est nécessaire, en formulant les questions d'évaluation, en mettant à disposition les ressources adéquates pour la réalisation de l'évaluation, en assurant la supervision des travaux, et en utilisant les résultats pour informer leur prise de décision. Les évaluateurs sont responsables des aspects techniques, à savoir la définition de la méthodologie, la constitution de l'échantillon d'évaluation, la collecte des données et l'analyse.

Concept clé :
Une évaluation est un partenariat entre des décideurs et des évaluateurs.

Une évaluation est un juste équilibre entre les compétences techniques et l'impartialité d'un groupe d'évaluateurs externes d'une part, et la pertinence politique, l'orientation stratégique et la coordination opérationnelle des décideurs d'autre part. Dans ce partenariat, le degré de séparation institutionnelle entre ceux qui réalisent l'évaluation et ceux qui en exploitent les résultats constitue un élément clé. L'indépendance des évaluateurs par rapport à l'institution responsable du projet qui fait l'objet de l'évaluation est primordiale pour en garantir l'objectivité. Toutefois, les évaluations peuvent souvent servir plusieurs objectifs, parmi lesquels le renforcement des capacités des institutions publiques en matière d'évaluation et la sensibilisation des gestionnaires du programme aux effets de leurs projets sur le terrain durant leur mise en œuvre.

Pour qu'une évaluation d'impact soit une réussite, les évaluateurs et les décideurs doivent impérativement collaborer. L'évaluation doit être menée par un groupe externe de manière à en assurer l'objectivité et la crédibilité ; toutefois, elle ne saurait être détachée des règles opérationnelles. Il convient en particulier de tenir compte des règles de mise en œuvre du programme pour garantir une bonne conception de l'évaluation et pour s'assurer que le programme et l'évaluation sont exécutés de manière coordonnée, l'un n'entravant pas l'autre. En outre, faute d'un engagement marqué des décideurs dès le début du processus, les résultats ont moins de chances d'avoir une pertinence politique directe ou d'influencer les politiques menées par les autorités.

Composition d'une équipe d'évaluation

Les décideurs peuvent mandater une évaluation d'impact sous diverses formes d'arrangements contractuels. Premièrement, l'institution publique commanditant l'évaluation peut décider de sous-traiter l'ensemble du travail. Elle doit alors établir au moins une version préliminaire du plan d'évaluation indiquant notamment les objectifs clés, les questions de politique, la méthodologie souhaitée, les données à collecter et les plafonds budgétaires. Ce plan fait office de cadre de référence pour lancer un appel d'offres techniques et financières auprès d'évaluateurs externes. Il peut également spécifier la composition minimum souhaitée de l'équipe d'évaluateurs externes. La préparation des propositions techniques est l'occasion pour les évaluateurs externes de suggérer des améliorations au plan d'évaluation établi par les autorités. Une fois l'évaluation contractée, l'agence externe retenue se charge de la gestion de l'évaluation et désigne un gestionnaire de l'évaluation. Dans ce cas de figure, les autorités se contentent d'un rôle de supervision.

Dans un deuxième type d'arrangement, l'institution publique qui commandite l'évaluation peut aussi décider d'en assurer la gestion directe. Dans ce cas, elle devra établir le plan d'évaluation et sous-traiter la réalisation de l'évaluation par composantes et par étapes successives. Le gestionnaire de l'évaluation est alors l'institution publique qui a demandé l'évaluation.

Indépendamment des dispositions contractuelles, l'une des principales tâches qui incombent au gestionnaire de l'évaluation est la constitution de l'équipe d'évaluation en tenant compte des intérêts des clients et des diverses étapes nécessaires pour mener l'évaluation à bien. Chaque évaluation est différente, mais l'équipe technique, qui doit assurer la collecte des données qualitatives et quantitatives, s'entourera dans presque tous les cas des personnes suivantes :

- *un gestionnaire de l'évaluation,* qui sera chargé de définir les objectifs clés, les questions de politique, les indicateurs et les besoins en matière d'informations (souvent en étroite collaboration avec les décideurs et à partir d'une théorie du changement comme la chaîne de résultats), de sélectionner la méthode d'évaluation, de constituer l'équipe d'évaluation et de préparer les termes de référence pour les composantes de l'évaluation qui seront sous-traitées. Il est important de choisir un gestionnaire de l'évaluation capable de travailler efficacement avec les organismes de collecte de données, les analystes et les décideurs qui utiliseront les données et les résultats de l'évaluation. Si le gestionnaire de l'évaluation n'est pas sur place, il est recommandé de désigner un gestionnaire local qui assurera la coordination du travail d'évaluation en collaboration avec le gestionnaire international.

- *un spécialiste en échantillonnage,* qui dirigera les travaux liés aux calculs de puissance et à l'échantillonnage. Pour les évaluations d'impact quantitatives, ce spécialiste doit effectuer les calculs de puissance pour déterminer la taille de l'échantillon adéquate selon les indicateurs retenus, sélectionner l'échantillon, analyser la validité de l'échantillon obtenu par rapport à l'échantillon prévu et formuler des conseils aux analystes en leur indiquant, le cas échéant, comment introduire des pondérations au moment de l'analyse. Cet expert pourra aussi sélectionner les sites ou groupes pour la phase pilote du projet. S'il s'agit d'un consultant international, il aura sans doute besoin d'être assisté d'un coordonnateur local qui collectera les données nécessaires au tirage de l'échantillon.

- *une personne ou une équipe responsable de la conception des instruments de collecte des données et des manuels les accompagnant,* qui veillera, en collaboration avec le gestionnaire de l'évaluation, à ce que ces instruments permettent bien de recueillir les données nécessaires à l'analyse et qui contribuera à l'essai des questionnaires durant la phase pilote.

- *une équipe de terrain,* qui comprendra, entre autres, un responsable de terrain chargé de la supervision de l'ensemble du travail de collecte des données, de la planification des opérations de collecte à la formation et à l'organisation des équipes de terrain, lesquelles sont généralement constituées de superviseurs et d'enquêteurs.
- *des gestionnaires de données et des agents de saisie,* qui devront concevoir les programmes de saisie des données, saisir et vérifier la validité des données, fournir la documentation nécessaire et produire des rapports présentant une description basique des données qui seront ensuite vérifiés par les analystes.
- *des analystes de données et des analystes stratégiques,* qui travailleront à partir des données fournies et en collaboration avec le gestionnaire de l'évaluation pour effectuer l'analyse et rédiger les rapports d'évaluation.

Partenaires de l'évaluation

L'une des premières questions sur laquelle les décideurs et le gestionnaire de l'évaluation doivent trancher est de savoir si l'évaluation (ou une partie de l'évaluation) peut être mise en œuvre localement et de déterminer le type de supervision et d'assistance extérieure nécessaires. Les capacités en matière d'évaluation varient beaucoup d'un pays à l'autre. Les contrats internationaux permettant à une société d'un pays donné de mener une évaluation dans un autre pays sont de plus en plus courants. Il est également de plus en plus fréquent que les gouvernements et les institutions internationales effectuent conjointement des évaluations au niveau local, tout en assurant une supervision internationale. C'est au gestionnaire de l'évaluation d'évaluer les capacités locales et de déterminer qui sera responsable des divers aspects de l'évaluation.

Une autre question qui se pose est de savoir s'il convient de travailler avec une société privée ou un organisme public. Les sociétés ou les instituts de recherche privés sont souvent plus à même de tenir le calendrier, mais dans ce cas, l'opportunité de renforcer les capacités dans le secteur public peut être perdue. En revanche, les sociétés privées sont parfois plus réticentes à intégrer des éléments qui rendront leurs efforts plus couteux. Les évaluations peuvent aussi être confiées à des instituts de recherche ou à des universités. La réputation et l'expertise technique de certains instituts de recherche ou de certaines universités peuvent constituer un gage de crédibilité des résultats obtenus et donc contribuer à leur acceptation immédiate par les parties prenantes au programme. Toutefois, ces organisations manquent parfois de l'expérience opérationnelle et des capacités nécessaires pour mener à bien certains aspects de l'évaluation, tels que la collecte des données. Ces aspects devront alors être confiés à d'autres partenaires. Dans tous les cas, quelle que soit la combinaison retenue, il est impératif d'étudier soigneusement l'expérience des éventuels collaborateurs en matière d'évaluation pour faire le bon choix.

En particulier, en considérant de travailler avec une institution publique, l'évaluateur doit bien étudier les capacités de l'équipe d'évaluation à la lumière des autres activités à sa charge. Ceci est encore plus vrai si l'institution en question assume des responsabilités multiples avec un personnel limité. Mieux vaut avoir une bonne idée de la charge de travail de l'institution afin d'évaluer si son volume de travail affectera la qualité de l'évaluation, mais aussi afin d'estimer le coût d'opportunité en termes d'autres tâches que l'institution pourrait réaliser à la place. Par exemple, une évaluation d'impact d'une réforme du système éducatif nécessitait la participation du personnel de l'équipe chargée de l'évaluation des examens nationaux semestriels. Cette équipe avait été associée à l'évaluation d'impact parce qu'elle regroupait les professionnels les plus qualifiés en la matière et que cette opération permettait une complémentarité entre l'évaluation d'impact et les examens nationaux. Toutefois, tant la réforme que l'évaluation d'impact durent être reportées. Ceci a non seulement remis en cause le travail d'enquête, mais a aussi retardé la réalisation des examens finaux qui n'ont pas eu lieu selon le calendrier prévu. En plus de l'évaluation, le pays a ainsi perdu une belle occasion de faire le suivi du progrès de son système éducatif. Il est possible d'éviter les problèmes de ce type en assurant une bonne coordination entre les responsables de l'unité chargée de l'évaluation d'impact, de manière à permettre une planification adéquate des diverses activités ainsi qu'une bonne répartition du personnel et des ressources.

Quand effectuer l'évaluation ?

Dans la première partie du présent manuel, nous avons évoqué les avantages des évaluations prospectives, prévues dès le début de la préparation du programme. Une planification précoce permet d'élargir les possibilités pour la constitution des groupes de comparaison, permet d'assurer la collecte des données de référence et contribue à établir un consensus sur les objectifs du programme et de l'évaluation entre les diverses parties prenantes.

Il est important de prévoir l'évaluation dès la phase de conception du projet, mais il peut être utile d'attendre que le projet ait acquis une certaine maturité avant de réaliser l'évaluation. Les projets pilotes ou les réformes nouvelles font souvent l'objet de révisions tant au niveau de leur contenu que de la manière, du moment, du lieu et des responsables de leur mise en œuvre. Les responsables du programme peuvent avoir besoin de temps pour intégrer et appliquer systématiquement de nouvelles règles opérationnelles. L'exercice d'évaluation exige que le programme soit mis en œuvre selon des règles opérationnelles précises pour pouvoir générer des contrefactuels adéquats. En ce sens, il est parfois préférable de réaliser des évaluations pour des programmes établis.

La collecte de *données de référence* est toujours nécessaire, mais la question du laps de temps requis avant de mesurer les résultats se pose souvent. Tout dépend du contexte : « Si l'évaluation a lieu trop tôt, il y a un risque de ne mesurer qu'un impact partiel ou nul ; si elle a lieu trop tard, il y a un risque que le programme ait perdu le soutien des donateurs ou des autorités ou qu'un mauvais programme ait

déjà été élargi » (King et Behrman 2009, p. 56). Les éléments suivants doivent être pris en considération lorsqu'il s'agit de déterminer le calendrier de collecte des données de suivi[4] :

- Le cycle du programme, notamment la durée, le temps nécessaire à la mise en œuvre et les retards éventuels
- Le temps jugé nécessaire pour que le programme produise des résultats ainsi que la nature des résultats à l'étude
- Les cycles d'élaboration des politiques publiques.

En premier lieu, l'évaluation d'impact doit être en adéquation avec le cycle de mise en œuvre du programme. L'évaluation ne doit pas modifier le plan de déroulement du programme. Par essence, l'évaluation est soumise au calendrier du programme ; elle doit se plier à la durée prévue du programme. Elle doit également s'adapter aux éventuels retards de mise en œuvre si les services prévus tardent à être offerts ou sont retardés par des facteurs externes[5]. En général, même s'il faut prévoir un calendrier d'évaluation dès la conception du programme, les évaluateurs doivent faire preuve de flexibilité et accepter de procéder à des modifications au fur et à mesure du déroulement du programme. Il faut en outre prévoir un bon système de suivi pour que le rythme de l'évaluation puisse s'adapter au rythme auquel les interventions se déroulent.

Le calendrier de *collecte des données de suivi* doit tenir compte du temps qui sera nécessaire après la mise en œuvre du programme pour que les résultats se matérialisent. La chaîne de résultats permet justement d'identifier les indicateurs de résultats et de définir le moment opportun pour les mesurer. Certains programmes (comme les programmes de filets sociaux) visent des bénéfices à court terme tandis que d'autres (comme les programmes d'éducation de base) sont plus orientés vers le long terme. De plus, certains résultats nécessitent, par nature, plus de temps pour se manifester (c'est le cas par exemple des résultats au plan de l'espérance de vie ou de la fécondité dans les réformes de santé) que d'autres (comme les programmes de formation).

Par exemple, dans le cadre de l'évaluation du Fonds d'investissement social en Bolivie, les données de référence ont été recueillies en 1993, mais il a fallu attendre jusqu'en 1998 pour collecter les données de suivi en raison du temps nécessaire pour que l'ensemble des interventions soient exécutées (projets d'approvisionnement en eau et de réseaux d'assainissement, cliniques et écoles) et pour que les effets sur l'éducation et la santé de la population se fassent sentir (Newman et al. 2002). Une période de temps similaire a été nécessaire pour l'évaluation du projet d'éducation primaire au Pakistan, qui reposait sur une approche expérimentale ayant recours à des données de référence et de suivi pour évaluer l'impact des écoles communautaires sur les résultats, notamment scolaires, des étudiants (King, Orazem et Paterno, 2008).

Le moment où la collecte des données de suivi doit avoir lieu dépend donc largement du programme et de l'indicateur des résultats à l'étude. Pour certaines évaluations, les données de suivi peuvent être recueillies alors que le programme est

en cours de mise en œuvre, ce qui permet de mesurer les impacts à court terme et de faire le suivi de l'échantillon d'évaluation de manière à limiter son attrition dans le temps. Pour les programmes dont les opérations sont limitées dans le temps, la collecte de données de suivi après la fin du programme peut permettre de mieux mesurer les changements à long terme. Des collectes de données de suivi peuvent même être organisées à plusieurs reprises, ce qui donne la possibilité d'analyser et de comparer les résultats à court et à moyen terme.

Les données de suivi collectées au cours de la mise en œuvre du programme peuvent ne pas suffire pour estimer l'impact total du programme si la mesure des indicateurs intervient trop tôt. En effet, « les programmes ne sont pas forcément pleinement efficaces au début de leur lancement. Les promoteurs et les bénéficiaires du programme ont besoin d'un temps d'apprentissage » (King et Behrman 2009, 65). Il n'en reste pas moins qu'il est très utile d'avoir des informations sur l'impact à court terme. Comme nous l'avons déjà souligné, certains programmes (comme les programmes de filets sociaux) visent principalement des objectifs à court terme. Des informations sur la performance à court terme d'un programme peuvent également donner des indications sur les résultats espérés à plus long terme. Les indicateurs à court terme permettent souvent de bonnes prédictions des indicateurs à plus long terme (par exemple, les naissances sous assistance médicale constituent un indicateur à court terme de l'évolution de la mortalité infantile). Les données de suivi collectées alors que le programme est en cours de mise en œuvre permettent aussi de dégager des résultats préliminaires de l'évaluation d'impact, ce qui peut être l'occasion de relancer le dialogue entre les évaluateurs et les décideurs.

Les données de suivi qui permettent de mesurer les résultats à long terme après la mise en œuvre du programme sont généralement celles qui permettent de cerner le mieux l'efficacité d'un programme. Par exemple, les résultats positifs mis en évidence par les évaluations de l'impact à long terme des programmes de développement de la petite enfance aux États-Unis (Currie et Thomas 1995, 2000 ; Currie 2001) et en Jamaïque (Grantham-McGregor et al. 1994) ont été déterminants dans la décision d'investir dans ces projets.

L'obtention d'impacts à long terme constitue parfois l'objectif explicite de certains programmes, mais ils peuvent aussi résulter d'effets imprévus et indirects, liés par exemple aux changements de comportement. La détermination de l'impact à long terme peut néanmoins se révéler problématique. L'impact peut tout simplement disparaître au fil du temps. Une méthodologie d'évaluation d'impact bien conçue peut être compromise. Par exemple, des effets de débordements peuvent se produire entre les bénéficiaires du programme et les unités du groupe de comparaison.

Bien que les données de suivi à court et à long terme soient complémentaires, le calendrier de l'évaluation doit tenir compte du moment opportun pour que les résultats de l'évaluation éclairent les prises de décision de politique publique. Il doit ainsi assurer la synchronisation des activités d'évaluation et de collecte de données avec les prises de décision majeures. La production des résultats doit être planifiée de manière à justifier les budgets, l'élargissement éventuel du programme ou toute autre décision stratégique de politique publique.

Comment établir le budget d'une évaluation d'impact ?

L'établissement du budget est l'une des dernières étapes dans la conception d'une évaluation d'impact. Dans cette section, nous allons examiner les coûts de certaines évaluations d'impact réalisées par le passé, aborder comment définir le budget d'une évaluation et suggérer quelques possibilités de financement.

Données sur les coûts

Les tableaux 10.2 et 10.3 présentent les coûts d'évaluations d'impact de quelques projets soutenus par la Banque mondiale. Les projets figurant dans le tableau 10.2 sont issus d'une revue exhaustive des programmes financés par l'unité Protection sociale et emploi. Ceux du tableau 10.3 ont été sélectionnés en fonction de la disponibilité des données budgétaires parmi les évaluations d'impact financées par le Fonds espagnol d'évaluation d'impact (SIEF). Ces deux échantillons ne sont pas nécessairement représentatifs de l'ensemble des évaluations menées par la Banque mondiale, d'autant plus que les données relatives aux coûts ne sont pas toujours disponibles, mais ils n'en constituent pas moins de bonnes références sur les coûts d'évaluations d'impact rigoureuses.

Tableau 10.2 Coûts d'évaluations d'impact de projets soutenus par la Banque mondiale

Évaluation d'impact (EI)	Pays	Coût total de l'EI (USD)	Coût total du programme (USD)	EI/coût total du programme (%)
Développement des compétences et de l'emploi des migrants	Chine	220 000	50 000 000	0,4
Projet de filet de protection sociale	Colombie	130 000	86 400 000	0,2
Programme d'investissement dans les secteurs sociaux	République dominicaine	600 000	19 400 000	3,1
Protection sociale	Jamaïque	800 000	40 000 000	2,0
Assistance technique Projet de filet de protection sociale	Pakistan	2 000 000	60 000 000	3,3
Projet de protection sociale	Panama	1 000 000	24 000 000	4,2
1er projet communautaire d'amélioration des conditions de vie	Rwanda	1 000 000	11 000 000	9,1
Phase 3 du projet de Fonds social pour le développement	Rép. du Yémen	2 000 000	15 000 000	13,3
Moyenne		**968 750**	**38 225 000**	**4,5**

Source : calculs des auteurs à partir d'un échantillon de programmes de la Banque mondiale dans le secteur de la protection sociale.
Remarque : EI = évaluation d'impact

Tableau 10.3 Répartition des coûts pour un échantillon de projets soutenus par la Banque mondiale

			Répartition des coûts de l'EI				
Évaluation d'impact du SIEF	**Pays**	**Coût total**	**Déplace-ments**	**Personnel Banque mondiale**	**Consultants (nationaux et internationaux)**	**Collecte de données (y.c. personnel)**	**Autres (coûts de diffusion et ateliers)**
Crédit d'appui à la réduction de la pauvreté et à la santé maternelle	Bénin	1 690 000	270 000	200 000	320 000	840 000	60 000
Rémunération à la performance des enseignants	Brésil	513 000	78 000	55 000	105 000	240 000	35 000
Programme Nadie es Perfecto pour améliorer les compétences parentales	Chili	313 000	11 500	—	35 500	260 000	6 000
Rémunération à la performance dans le secteur de la santé : évaluation du projet Santé XI	Chine	308 900	60 000	35 000	61 000	152 900	—
Programme national de garantie de l'emploi rural	Inde	390 000	41 500	50 000	13 500	270 000	15 000
Éducation, Santé et Nutrition/ Rôle du contrôle du paludisme dans l'amélioration de l'éducation	Kenya	652 087	69 550	60 000	103 180	354 000	65 357
Campagne de prévention du sida chez les jeunes : abstinence, fidélité et sexualité sans risque	Lesotho	630 300	74 300	9 600	98 400	440 000	8 000
TMC, scolarisation et risque de sida	Malawi	1 842 841	83 077	144 000	256 344	1 359 420	—
Programme ContigoVamos por Mas Oportunidades dans l'État de Guanajuato	Mexique	132 199	2 660	50 409	—	80 640	1 150

Projet pilote TMC et éducation en milieu rural	Maroc	674 367	39 907	66 000	142 460	426 000	—
Apprendre et grandir avec le VIH/sida : assignation aléatoire d'un programme de développement de la petite enfance	Mozambique	838 650	86 400	31 000	62 500	638 750	20 000
Formation des distributeurs communautaires à la prévention et au traitement du paludisme	Nigéria	1 024 040	64 000	35 000	106 900	817 740	—
Éducation, Santé et Nutrition/ Rôle du contrôle du paludisme dans l'amélioration de l'éducation	Sénégal	644 047	61 800	60 000	102 890	354 000	65 357
Les TMC pour éviter le sida et d'autres maladies sexuellement transmissibles	Tanzanie	771 610	60 000	62 000	100 000	518 611	30 999
Moyenne		**744 646**	**71 621**	**66 031**	**115 975**	**482 290**	**30 686**

Source : calculs des auteurs à partir d'un échantillon d'évaluations d'impact financées par le Fonds espagnol d'évaluation d'impact.

Remarque : TMC = transferts monétaires conditionnels ;— = non disponible ; SIEF, Fonds espagnol d'évaluation d'impact (Spanish Impact Evaluation Fund).

Les coûts directs des activités d'évaluation vont de 130 000 à deux millions de dollars pour un coût moyen de 968 750 dollars. Ces coûts sont très variables d'une évaluation à l'autre et peuvent paraître élevés en valeur absolue. Toutefois, en termes relatifs, ils ne dépassent pas 4,5 % en moyenne (fourchette comprise entre 0,2 % et 13,3 %) du coût total du programme[6]. À partir de l'échantillon de projets étudié, il apparaît que les évaluations d'impact ne représentent qu'un pourcentage limité du budget total d'un programme. Il convient en outre de comparer les coûts de l'évaluation d'impact au coût d'opportunités en l'absence d'une évaluation rigoureuse et, par conséquent, au risque de mise en œuvre d'un programme inefficace. Les évaluations permettent aux chercheurs et aux décideurs d'identifier les programmes ou les composantes d'un programme qui fonctionnent et ceux qui ne fonctionnent pas, et de déterminer les stratégies les plus efficaces pour atteindre les objectifs du programme. Dans cette perspective, les ressources nécessaires à la réalisation d'une évaluation d'impact constituent un investissement relativement faible au vu de l'utilité d'un tel travail.

Le tableau 10.3 présente la répartition des coûts d'un échantillon d'évaluations d'impact financées par le Fonds espagnol d'évaluation d'impact (SIEF). Le coût total englobe le temps de travail du personnel de la Banque mondiale et des consultants nationaux et internationaux, les déplacements, la collecte des données et les activités de diffusion de l'information[7]. Dans les évaluations figurant dans le tableau, comme dans presque toutes les évaluations où les données existantes ne peuvent pas être utilisées, ce sont les coûts de collecte des données qui sont les plus importants : ils ne représentent pas moins de 60 % du coût total en moyenne.

Il est important de souligner que ces chiffres concernent des évaluations de taille et de type différents. Le coût relatif de l'évaluation d'un programme pilote est généralement plus élevé que celui d'un programme d'envergure nationale ou ouvert à l'ensemble de la population. De plus, certaines évaluations ne nécessitent qu'une enquête de suivi ou peuvent se fonder sur des données existantes, tandis que d'autres exigent plusieurs opérations de collecte de données. Le manuel sur les Enquêtes sur le niveau de vie des ménages[1] (Grosh et Glewwe, 2000) donne une estimation des coûts des opérations de collecte d'enquête de ménages dans divers pays. Les auteurs de l'étude insistent sur le fait que les coûts encourus dépendent largement des capacités de l'équipe locale, des ressources disponibles et du temps passé sur le terrain. Pour réaliser une meilleure estimation des coûts d'une enquête dans un contexte donné, il est recommandé de commencer par contacter les services statistiques nationaux.

Estimation du budget d'une évaluation d'impact

Il est évident que de nombreuses ressources doivent être mobilisées pour réaliser une évaluation d'impact. Le budget comprend les frais de personnel pour, au minimum, un chercheur, un assistant de recherche, un coordinateur de terrain, un spécialiste de l'échantillonnage, des enquêteurs et le personnel du projet qui peut

apporter un appui dans le cadre de l'évaluation. Ces ressources humaines peuvent aussi comprendre des chercheurs et des experts d'organisations internationales, des consultants locaux ou internationaux et du personnel local travaillant pour le programme. Aux frais de personnel s'ajoutent les frais de déplacements et de mission (hôtels et indemnités quotidiennes) ainsi que les frais de diffusion, souvent sous forme d'ateliers, de rapports et de publications académiques.

Comme nous l'avons déjà souligné, les coûts les plus importants d'une évaluation sont les coûts relatifs à la collecte des données (y compris la création et la mise en œuvre d'une enquête pilote), au matériel et aux équipements nécessaires à cette collecte, à la formation et au salaire journalier des enquêteurs, aux véhicules et à l'essence ainsi qu'aux opérations de saisie des données. Pour calculer le coût de ces intrants, il est nécessaire de faire quelques hypothèses sur, par exemple, le temps nécessaire pour réaliser un questionnaire ou le temps de déplacement entre les sites. Le tableau 10.4 présente une feuille de calcul permettant d'estimer les coûts de la collecte des données.

Les coûts d'une évaluation d'impact peuvent être répartis sur plusieurs exercices. Le tableau 10.5 montre comment les coûts de chaque étape d'une évaluation peuvent être répartis sur plusieurs exercices à des fins comptables et de reporting. Les besoins financiers sont plus élevés les années où une collecte de données est réalisée.

Financement des évaluations d'impact

Une évaluation d'impact peut être financée à partir de plusieurs sources, dont les prêts-projets, les budgets directs des programmes, les subventions de recherche ou le financement de donateurs. Les équipes d'évaluation se tournent souvent vers plusieurs sources pour réunir les fonds nécessaires. Les évaluations ont été traditionnellement principalement financées par des budgets de recherche, mais les sources de financement se diversifient de plus en plus avec le développement croissant des pratiques d'élaboration des politiques fondée sur les preuves. Lorsque l'enjeu d'un programme est important pour l'ensemble d'une communauté et qu'une évaluation solide et crédible peut être mise en place pour acquérir de nouvelles connaissances, les décideurs doivent être encouragés à rechercher des financements extérieurs, particulièrement puisque les résultats de l'évaluation constituent un bien public. Parmi les financeurs potentiels figurent l'État, les banques de développement, les organisations multilatérales, les organismes des Nations Unies, les fondations, les mécènes ainsi que les instituts de recherche et d'évaluation tels que l'Initiative internationale pour l'évaluation d'impact.

Tableau 10.4 Feuille de calcul pour l'estimation du coût d'une évaluation d'impact

Tâches et ressources	Nombre	Taux/ unité	Nombre d'unités	Total
Personnel				
Personnel chargé de l'évaluation (gestionnaire de l'évaluation, etc.)				
Consultants internationaux et/ou nationaux (chercheurs/responsable d'enquêtes)				
Assistant de recherche				
Statisticien				
Coordonnateur de terrain				
Déplacements				
Billets d'avion/voyages nationaux et internationaux				
Transports routiers				
Frais de mission (hôtels et indemnités journalières)				
Collecte de données[a]				
Conception de l'instrument				
Pilotage				
Formation				
Déplacements et indemnités journalières				
Matériel et équipement pour l'enquête				
Impression des questionnaires				
Personnel de terrain				
Enquêteurs				
Superviseurs				
Transport (véhicules et essence)				
Chauffeurs				
Saisie et nettoyage des données				
Analyse et diffusion des données				
Ateliers				
Articles, rapports				
Autres				
Bureaux				
Communications				
Logiciels				

a. Les calculs relatifs à la collecte de données doivent refléter les hypothèses telles que le nombre de rondes de collecte nécessaires, le temps nécessaire à la collecte, le nombre de villages dans l'échantillon, le nombre de ménages par village, la longueur du questionnaire, les temps de déplacement, etc.

Tableau 10.5 Budget d'une évaluation d'impact

	Phase de conception				Phase de collecte des données de référence			
	Unités	**Coût par unité (USD)**	**Nombre d'unités**	**Coût total (USD)**	**Unités**	**Coût par unité (USD)**	**Nombre d'unités**	**Coût total (USD)**
A. Salaires du personnel	Semaines	7 500	2	15 000	Semaines	7 500	2	15 000
B. Frais de consultants				10 250				27 940
Consultant international (1)	Jours	450	15	6 750	Jours	450	0	0
Consultant international (2)	Jours	350	10	3 500	Jours	350	10	3 500
Assistant de recherche/coordonnateur de terrain	Jours	188	0	0	Jours	188	130	24 440
C. Déplacements et frais de mission				14 100				15 450
Personnel : billets d'avion internationaux	Voyages	3 350	1	3 350	Voyages	3 350	1	3 350
Personnel : hôtels et indemnités quotidiennes	Jours	150	5	750	Jours	150	5	750
Billets internationaux : consultants internationaux	Voyages	3 500	2	7000	Voyages	3 500	2	7000
Hôtels et indemnités quotidiennes : consultants internationaux	Jours	150	20	3 000	Jours	150	20	3 000
Billets internationaux : coordonnateur de terrain	Voyages		0	0	Voyages	1 350	1	1 350
Hôtels et indemnités quotidiennes : coordonnateur de terrain	Jours		0	0	Jours	150	0	0
D. Collecte de données								126 000
Données type 1 : consentement					École	120	100	12 000
Données type 2 : résultats volet Éducation					Enfant	14	3 000	42 000
Données type 3 : résultats volet Santé					Enfant	24	3 000	7 200
V. Autres								
Atelier(s)								
Diffusion/reporting								
Autres 1 (frais généraux de coordination)								
Coût total par phase	Phase de conception			39 350	Phase de collecte des données de référence			184 390

(à suivre)

Tableau 10.5 *(suite)*

	Données de suivi Phase I				Données de suivi Phase II			
	Unités	**Coût par unité (USD)**	**Nombre d'unités**	**Coût total (USD)**	**Unités**	**Coût par unité (USD)**	**Nombre d'unités**	**Coût total (USD)**
A. Salaires du personnel	Semaines	7 500	2	15 000	Semaines	7 500	2	15 000
B. Frais de consultants				32 550				32 440
Consultant international (1)	Jours	450	15	6 750	Jours	450	10	4 500
Consultant international (2)	Jours	350	20	7 000	Jours	350	10	3 500
Assistant de recherche/coordonnateur de terrain	Jours	188	100	18 800	Jours	188	130	24 440
C. Déplacements et frais de mission				20 000				20 000
Personnel : billets d'avion internationaux	Voyages	3 350	2	6 700	Voyages	3 350	2	6 700
Personnel : hôtels et indemnités quotidiennes	Jours	150	10	1 500	Jours	150	10	1 500
Billets internationaux : consultants internationaux	Voyages	3 500	2	7 000	Voyages	3 500	2	7 000
Hôtels et indemnités quotidiennes : consultants internationaux	Jours	150	20	3 000	Jours	150	20	3 000
Billets internationaux : coordonnateur de terrain	Voyages	1 350	1	1 350	Voyages	1 350	1	1 350
Hôtels et indemnités quotidiennes : coordonnateur de terrain	Jours	150	3	450	Jours	150	3	450
D. Collecte de données				114 000				114 000
Données type 1 : consentement								
Données type 2 : résultats volet Éducation	Enfant	14	3 000	42 000	Enfant	14	3 000	42 000
Données type 3 : résultats volet Santé	Enfant	24	3 000	72 000	Enfant	24	3 000	72 000
V. Autres								**65 357**
Atelier(s)						20 000	2	40 000
Diffusion/reporting						5 000	3	15 000
Autres 1 (frais généraux de coordination)						5 179	2	10 357
Coût total pas phase	Phase de suivi I			181 550	Phase de suivi II			246 797
							Coût total de l'évaluation :	652 087

Notes

1. Le contenu de cette section s'applique plus directement à la méthode de l'assignation aléatoire, mais les mêmes principes s'appliquent aux évaluations basées sur d'autres méthodes.
2. Voir Kimmel 1988 ; NIH 2006 ; USAID 2008 ; U.S. Department of Health and Human Services 2010 ; et U.S. National Archives 2009.
3. Parmi les risques et difficultés associés à la collecte de données pour l'évaluation de programmes sociaux citons l'impossibilité d'obtenir le consentement éclairé des sujets, l'évaluation du développement cognitif des enfants en présence des parents qui peut donner lieu à des suppositions sur leur développement futur, le fait de demander à parler en privé à des femmes ou d'interviewer des femmes sur des sujets sensibles en présence d'hommes de la famille, le fait d'ignorer le temps ou coût d'opportunité de participer à une enquête et l'offre d'une compensation le cas échéant.
4. Pour de plus amples détails sur les questions de calendrier des évaluations de programmes sociaux, voir King et Behrman (2009).
5. « Plusieurs raisons peuvent expliquer pourquoi la mise en œuvre d'un programme n'est pas immédiate ou parfaite, pourquoi la durée d'exposition à un traitement varie non seulement d'une zone à l'autre, mais aussi entre chaque bénéficiaire final, et pourquoi des temps d'expositions différents peuvent conduire à l'estimation d'impacts différents » (King et Behrman 2009, 56).
6. Dans ce cas, le coût est exprimé en pourcentage de la part du coût du projet financée par la Banque mondiale.
7. Ce chiffre ne comprend pas les coûts du personnel local souvent très impliqué dans la conception et la supervision de l'évaluation, car les données relatives à ces coûts sont rarement disponibles.

Références

Behrman, Jere R. et John Hoddinott. 2001. « An Evaluation of the Impact of PROGRESA on Pre-school Child Height. » FCND Briefs 104, International Food Policy Research Institute, Washington, DC.

Currie, Janet. 2001. « Early Childhood Education Programs. » *Journal of Economic Perspectives* 15 (2) : 213–38.

Currie, Janet et Duncan Thomas. 1995. « Does Head Start Make a Difference ? » *American Economic Review* 85 (3) : 34164.

——. 2000. « School Quality and the Longer-Term Effects of Head Start. » *Journal of Economic Resources* 35 (4) : 75574.

Gertler, Paul J. 2004. « Do Conditional Cash Transfers Improve Child Health ? Evidence from PROGRESA's Control Randomized Experiment. » *American Economic Review* 94 (2) : 33641.

Grantham-McGregor, S., C. Powell, S. Walker et J. Himes. 1994. « The Long-Term Follow-up of Severely Malnourished Children Who Participated in an Intervention Program. » *Child Development* 65 : 428–93.

Grosh, Margaret et Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries : Lessons from 15 Years of the Living Standards Measurement Study*, vols. 1, 2 et 3. Washington DC : Banque mondiale.

Grosh, Margaret, Carlo del Ninno, Emil Tesliuc et Azedine Ouerghi. 2008. *For Protection and Promotion : The Design and Implementation of Effective Safety Nets*. Washington DC : Banque mondiale.

Jalan, Jyotsna et Martin Ravallion. 2003a. « Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. » *Journal of Business & Economic Statistics* 21 (1) : 19–30.

———. 2003b. « Does Piped Water Reduce Diarrhea for Children in Rural India ? » *Journal of Econometrics* 112 (1) : 15373.

Kimmel, Allan. 1988. *Ethics and Values in Applied Social Research*. Californie : Sage Publications.

King, Elizabeth M. et Jere R. Behrman. 2009. « Timing and Duration of Exposure in Evaluations of Social Programs. » *World Bank Research Observer* 24 (1) :55–82.

King, Elizabeth M., Peter F. Orazem et Elizabeth M. Paterno. 2008. « Promotion with and without Learning : Effects on Student Enrollment and Dropout Behavior. » Document de travail consacré à la recherche sur les politiques 4722, Banque mondiale, Washington, DC.

Levy, Santiago et Evelyne Rodríguez. 2005. *Sin Herencia de Pobreza : El Programa Progresa-Oportunidades de México*. Washington DC : Banque interaméricaine de développement.

NIH (U.S. National Institutes of Health). 2006. « Regulations and Ethical Guidelines » et « Rapport Belmont ». Office of Human Subjects Research. http://ohsr.od.nih.gov/index.html.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa et Jose Luis Evia. 2002. « An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund. » *Étude économique de la Banque mondiale* 16 (2) : 241–74.

Rosenbaum, Paul. 2002. *Observational Studies*. Springer Series in Statistics.

Rosenbaum, Paul et Donald Rubin. 1983. « The Central Role of the Propensity Score in Observational Studies of Causal Effects. » *Biometrika* 70 (1) : 41–55.

Schultz, Paul. 2004. « School Subsidies for the Poor : Evaluating the Mexican Progresa Poverty Program. » *Journal of Development Economics* 74 (1) : 199–250.

Skoufias, Emmanuel et Bonnie McClafferty. 2001. « Is Progresa Working ? Summary of the Results of an Evaluation by IFPRI. » Institut international de recherche sur les politiques alimentaires, Washington, DC.

USAID (agence américaine pour le développement international). 2008. « Procedures for Protection of Human Subjects in Research Supported by USAID. » http://www.usaid.gov/policy/ads/200/humansub.pdf.

U.S. Department of Health and Human Services. 2010. « International Compilation of Human Research Protections. » Office for Human Research Protections. http://www.hhs.gov/ohrp/international/HSPCompilation.pdf.

U.S. National Archives. 2009. « Protection of Human Subjects. » *U.S. Code of Federal Regulations*, Titre 22, partie 225.

CHAPITRE 11

Choisir l'échantillon

Une fois que vous avez choisi une méthode de sélection du groupe de comparaison, l'étape suivante de la planification d'une évaluation d'impact consiste à déterminer les données et l'échantillon nécessaires pour estimer avec précision les différences de résultats entre le groupe de traitement et le groupe de comparaison. Vous devez déterminer la taille de l'échantillon et la façon de prélever les unités de la population à l'étude pour former cet échantillon.

Quelles sont les données nécessaires ?

Il est essentiel de disposer de données de qualité pour évaluer l'impact de l'intervention sur les résultats à l'étude. La chaîne de résultats abordée au chapitre 2 constitue un bon point de départ pour définir les indicateurs à mesurer et le moment le plus propice pour le faire. Les données les plus essentielles sont celles qui permettent de mesurer les indicateurs de résultats directement affectés par le programme. L'évaluation d'impact ne doit toutefois pas se réduire à la mesure des résultats que le programme vise directement. Des données sur des indicateurs de résultats indirectement affectés par le programme ou sur des indicateurs reflétant les effets involontaires du programme augmentent la valeur des informations générées par l'évaluation d'impact. Comme nous l'avons vu au chapitre 2, les indicateurs de résultats doivent de préférence être spécifiques, mesurables, attribuables, réalistes et ciblés.

Les évaluations d'impact sont généralement réalisées sur plusieurs périodes, et vous devez donc déterminer le moment adéquat pour mesurer les indicateurs de

résultats. En suivant la chaîne de résultats, vous pouvez établir un classement des indicateurs de résultats allant des indicateurs à court terme (comme les taux de scolarisation dans le contexte d'un programme éducatif) aux indicateurs à long terme (comme l'achèvement des études ou l'insertion professionnelle). Afin de mesurer l'impact de manière fiable au fil du temps, des données sur ces indicateurs doivent dans la mesure du possible être collectées dès l'enquête de référence. La section du chapitre 10 consacrée au calendrier des évaluations apporte des indications sur le moment le plus propice pour collecter les données de suivi.

Concept clé :
Les indicateurs choisis doivent couvrir toute la chaîne de résultats afin de mesurer les résultats finaux, les résultats intermédiaires, la mise en œuvre de l'intervention, les facteurs exogènes et les caractéristiques de contrôle.

Nous allons voir que certains indicateurs peuvent ne pas se prêter à une évaluation d'impact si les échantillons sont de taille réduite. En effet, la taille des échantillons nécessaire pour mesurer les impacts sur des indicateurs de résultats extrêmement variables, rares ou susceptibles de n'être que légèrement affectés par une intervention, peut être prohibitive. Par exemple, pour cerner l'impact d'une intervention sur le taux de mortalité maternelle, un échantillon doit contenir un grand nombre de femmes enceintes. Dans ce cas, il peut être utile d'axer l'évaluation d'impact sur des indicateurs pour lesquels il existe une puissance suffisante pour détecter un impact.

Outre les indicateurs de résultat, il est également utile de prendre en compte les éléments suivants :

- *Données administratives sur la mise en œuvre de l'intervention.* Il faut au moins disposer de données de suivi pour savoir quand un programme débute et qui en bénéficie ainsi que pour pouvoir mesurer l'intensité de l'intervention dans les cas où tous les bénéficiaires ne bénéficient pas du même contenu, de la même qualité ou de la même durée de traitement.

- *Données sur les facteurs exogènes susceptibles d'influer sur le résultat à l'étude.* Ces données permettent de vérifier s'il existe des influences extérieures. Cet aspect est particulièrement important lors de l'utilisation de méthodes d'évaluation reposant sur un plus grand nombre d'hypothèses que les méthodes aléatoires. La prise en compte de variables de contrôle permet également de renforcer la puissance statistique.

- *Données sur d'autres caractéristiques.* L'inclusion de variables de contrôles supplémentaires ou l'analyse de l'hétérogénéité des effets du programme selon certaines caractéristiques permet d'affiner l'estimation des effets du traitement.

En résumé, il est nécessaire d'obtenir des indicateurs tout au long de la chaîne de résultats, y compris des indicateurs de résultats finaux, des indicateurs de résultats intermédiaires, et des mesures de la mise en œuvre de l'intervention, des facteurs exogènes et des caractéristiques de contrôl[1].

La méthodologie d'évaluation d'impact choisie détermine aussi les données nécessaires. Par exemple, si vous choisissez la méthode de l'appariement ou de la double différence, il vous faudra collecter des données portant sur une large gamme de caractéristiques à la fois pour le groupe de traitement et pour le groupe de comparaison, par exemple pour pouvoir effectuer les tests de robustesse décrits dans la deuxième partie du manuel.

Il est utile d'élaborer, pour chaque évaluation, une matrice comprenant la liste des questions à l'étude, les indicateurs de résultats pour chaque question, les autres types d'indicateurs indispensables et les sources des données, comme indiqué à la figure 2.3 (chapitre 2).

Les données existantes sont-elles suffisantes ?

Certaines données existantes sont presque toujours indispensables au début d'un programme pour estimer les valeurs de référence des indicateurs ou pour effectuer des calculs de puissance, comme nous le verrons plus loin. Au-delà de l'étape de planification, l'utilisation de données existantes peut nettement diminuer le coût d'une évaluation d'impact.

Toutefois, il est rare que les données existantes suffisent. Les évaluations d'impact nécessitent des données exhaustives couvrant un échantillon suffisamment important et représentatif à la fois du groupe de traitement et du groupe de comparaison. Des *données de recensement* couvrant l'ensemble des groupes de traitement et de comparaison sont rarement disponibles. Même si des recensements ont été réalisés, les données ne contiennent généralement qu'un nombre limité de variables ou ne sont pas collectées régulièrement. Les enquêtes nationales auprès des ménages comportent parfois une gamme étendue de variables, mais contiennent rarement suffisamment d'observations à la fois pour le groupe de traitement et le groupe de comparaison pour permettre une évaluation d'impact. Admettons par exemple que vous souhaitiez évaluer un vaste programme national qui concerne 10 % des ménages dans un pays donné. Si une enquête nationale est réalisée chaque année auprès de 5 000 ménages, elle couvrira peut-être 500 ménages bénéficiant du programme évalué. Cet échantillon est-il suffisant pour réaliser une évaluation d'impact ? Les calculs de puissance peuvent permettre de répondre à cette question, mais dans la plupart des cas, la réponse est non.

Il convient toutefois d'envisager sérieusement l'utilisation de *données administratives* existantes pour réaliser des évaluations d'impact. Les données administratives sont collectées dans le cadre des activités ordinaires des organismes responsables de l'exécution des programmes, le plus souvent au moment de la prestation des services. Dans certains cas, les données de suivi contiennent des indicateurs de résultats. Par exemple, certaines écoles compilent les taux de scolarisation, de fréquentation ou les résultats aux examens tandis que certains centres de santé enregistrent les données anthropométriques et les vaccinations ou les dossiers de santé de leurs patients. Certaines évaluations rétrospectives marquantes sont fondées sur des registres administratifs (par exemple, l'étude de Galiani, Gertler et Schargrodsky publiée en 2005 sur la politique d'alimentation en eau en Argentine).

Afin de déterminer si les données existantes peuvent être utilisées pour une évaluation d'impact donnée, les questions suivantes doivent être posées :

- *Taille.* Les bases de données existantes sont-elles assez grandes pour détecter un changement des indicateurs de résultats avec une puissance suffisante ?

- *Échantillonnage.* Les données existantes sont-elles disponibles à la fois pour le groupe de traitement et le groupe de comparaison ? Les échantillons existants sont-ils prélevés à partir d'un cadre d'échantillonnage correspondant à la population à l'étude ? Les unités ont-elles été prélevées du cadre d'échantillonnage à l'aide d'une méthode probabiliste ?

- *Portée.* Les données existantes contiennent-elles tous les indicateurs nécessaires pour répondre à toutes les questions de politique à l'étude ?

- *Fréquence.* La collecte des données existantes est-elle suffisamment fréquente ? Des données existantes sont-elles à disposition pour toutes les unités de l'échantillon et toute la période étudiée ?

Il est relativement rare que des données existantes soient suffisantes pour une évaluation d'impact. Vous devrez par conséquent fort probablement prévoir un budget pour la collecte de nouvelles données. La collecte des données représente souvent un coût important, mais il s'agit également d'un investissement à rendement élevé dont dépend la qualité de l'évaluation.

Dans certains cas, les données nécessaires à l'évaluation d'impact peuvent être collectées en déployant de nouveaux systèmes d'information, pour autant que ce déploiement soit conforme à la méthodologie d'évaluation adoptée, en particulier que les indicateurs de résultats soient collectés pour le groupe de traitement et le groupe de comparaison. Il peut être nécessaire de lancer de nouveaux systèmes d'information avant le lancement de nouvelles interventions afin que les centres administratifs du groupe de comparaison utilisent le nouveau système avant de recevoir l'intervention à évaluer. Étant donné que la qualité des données administratives peut varier, des audits et des vérifications externes sont nécessaires pour garantir la fiabilité de l'évaluation. La collecte de données d'évaluation d'impact par le biais de sources administratives au lieu d'enquêtes peut nettement réduire le coût de l'évaluation, mais n'est pas toujours faisable.

Si les données administratives ne sont pas suffisantes pour votre évaluation, vous devrez avoir recours à des *données d'enquête*. Il vous faudra alors déterminer si vous pouvez utiliser les enquêtes existantes ou si de nouvelles initiatives nationales de collecte de données sont prévues (par exemple des enquêtes démographiques, sanitaires ou de mesure des niveaux de vie des ménages). Si une enquête couvrant les indicateurs à l'étude est prévue, il peut être possible d'étendre l'échantillonnage pour les besoins de l'évaluation. Par exemple, l'évaluation du Fonds social du Nicaragua repose sur les données d'une enquête nationale sur la mesure du niveau de vie des ménages complété d'un échantillon supplémentaire de bénéficiaires (Pradhan et Rawlings 2002). Si une enquête prévue couvre la population à l'étude, il peut être possible d'y ajouter une série de questions aux fins de l'évaluation.

La plupart des évaluations d'impact nécessitent la collecte de données d'enquêtes, dont au moins une *enquête de référence* et une *enquête de suivi*. Les données d'enquête peuvent être de différents types en fonction du programme à évaluer et des unités analysées. La plupart des évaluations prennent pour principale source de données des enquêtes réalisées auprès de personnes ou de ménages. Nous allons nous attacher ci-dessous aux principes généraux de collecte des données

d'enquête. S'ils s'appliquent principalement aux enquêtes auprès des ménages, ces principes peuvent également être appliqués à la plupart des autres types de données d'enquête[2].

Avant de décider si vous allez utiliser les données existantes ou collecter de nouvelles données d'enquête, il convient de déterminer la taille de l'échantillon nécessaire. Si les données existantes contiennent un nombre suffisant d'observations, vous pourriez être en mesure de les utiliser. Dans le cas contraire, des données supplémentaires devront être collectées. Une fois que vous avez décidé de collecter des données d'enquête pour votre évaluation, vous devez :

- déterminer qui va collecter les données ;
- élaborer et tester des questionnaires ;
- effectuer des travaux de collecte sur le terrain et des contrôles de qualité ; et
- traiter et stocker les données.

Dans la suite de ce chapitre, nous verrons comment déterminer la taille de l'échantillon nécessaire et la façon de procéder à l'échantillonnage. Les autres étapes de la collecte de données sont abordées au chapitre 12. La mise en œuvre des différentes étapes est généralement confiée à un organisme indépendant, mais il est essentiel de comprendre leur portée et leurs principales composantes pour gérer efficacement une évaluation d'impact.

Calculs de puissance : quelle est la taille de l'échantillon nécessaire ?

Au moment de s'interroger sur l'éventuelle utilisation de données existantes ou la collecte de nouvelles données, la première étape consiste à définir la taille de l'échantillon nécessaire. Les calculs effectués dans ce but sont appelés des *« calculs de puissance »*. Dans cette section, nous évoquerons l'intuition sous-jacente aux calculs de puissance en nous concentrant sur le cas le plus simple : une évaluation réalisée à l'aide de la méthode de l'assignation aléatoire, en partant du principe que l'adhérence est totale. (L'adhérence totale signifie que toutes les unités assignées au groupe de traitement reçoivent effectivement le traitement et que toutes celles qui sont assignées au groupe de comparaison ne le reçoivent effectivement pas.)

Concept clé :
Les calculs de puissance indiquent la taille de l'échantillon nécessaire pour qu'une évaluation fournisse une estimation précise de l'impact d'un programme (c'est-à-dire de la différence des résultats entre le groupe de traitement et le groupe de comparaison).

Objectifs des calculs de puissance

Les calculs de puissance indiquent la taille minimum de l'échantillon nécessaire pour réaliser une évaluation d'impact et pour répondre de manière fiable à la question de politique à l'étude. Ils peuvent notamment être utilisés pour :

- Considérer si les bases de données existantes sont assez grandes pour réaliser une évaluation d'impact.

- Éviter de collecter trop d'informations, ce qui peut s'avérer très coûteux.
- Éviter de collecter trop peu de données. Admettons que vous souhaitiez évaluer un programme qui a un impact positif sur ses bénéficiaires. Si l'échantillon est trop petit, vous risquez de ne pas pouvoir détecter cet impact positif et de conclure que le programme n'a pas eu d'effet. Ceci peut amener les décideurs à supprimer le programme, dans ce cas au détriment des bénéficiaires et de la société.

Les calculs de puissance indiquent la taille minimale de l'échantillon (et donc du budget minimal) nécessaire pour mesurer l'impact d'un programme, à savoir le plus petit échantillon permettant d'identifier des différences pertinentes de résultats entre le groupe de traitement et le groupe de comparaison. Les calculs de puissance sont essentiels pour correctement déterminer les programmes qui fonctionnent et ceux qui ne fonctionnent pas.

L'impact du programme est-il différent de zéro ?

La plupart des évaluations d'impact cherchent à tester une hypothèse simple qui se résume par la question suivante : *le programme a-t-il un impact ?* Autrement dit, *l'impact du programme est-il différent de zéro ?* Deux étapes sont nécessaires pour répondre à cette question :

1. Estimer les résultats moyens pour le groupe de traitement et pour le groupe de comparaison.
2. Déterminer s'il existe une différence entre le résultat moyen du groupe de traitement et celui du groupe de comparaison.

Estimer les résultats moyens du groupe de traitement et du groupe de comparaison

Admettons que vous souhaitiez estimer l'impact d'un programme de nutrition sur le poids des enfants de cinq ans. Nous partons de l'hypothèse selon laquelle 100 000 enfants ont participé au programme et 100 000 enfants n'y ont pas participé, les participants ayant été sélectionnés de manière aléatoire parmi les 200 000 enfants du pays. Dans un premier temps, vous devrez estimer le poids moyen des participants et des non participants.

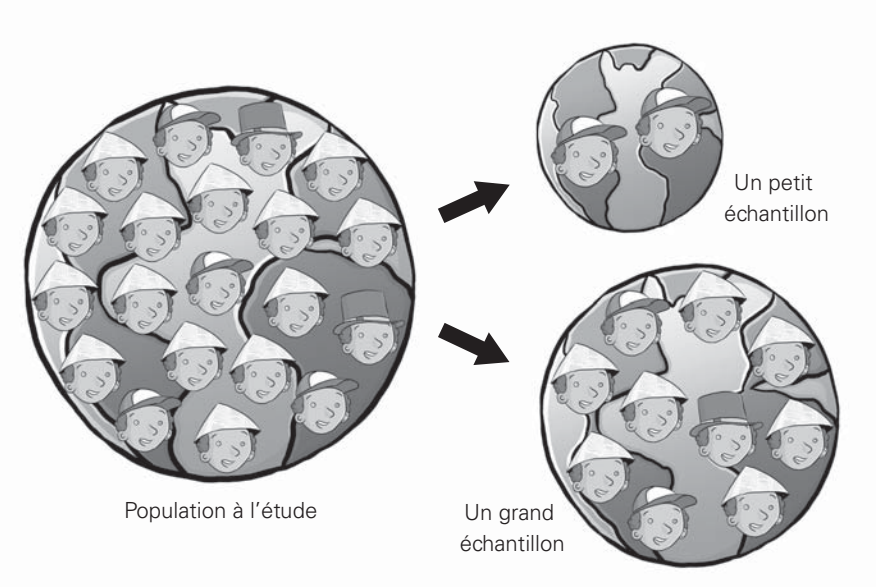
Pour déterminer le poids moyen des enfants participant[3] au programme, vous pourriez peser chacun d'entre eux puis calculer la moyenne. Cette approche serait évidemment extrêmement coûteuse. Heureusement, il n'est pas nécessaire de peser chaque enfant. En effet, la moyenne peut être estimée à partir du poids moyen d'un échantillon prélevé sur la population d'enfants participants[4]. Plus l'échantillon est grand, plus la moyenne de l'échantillon se rapprochera de la moyenne réelle. Si l'échantillon est petit (deux enfants par exemple), le poids moyen constituera une estimation très imprécise de la moyenne pour la population à l'étude. En revanche, un échantillon de 10 000 enfants donnera une estimation plus précise et plus proche

du véritable poids moyen. De manière générale, plus le nombre d'observations dans un échantillon est élevé, plus les statistiques qui en sont extraites sont fiables[5].

La figure 11.1 illustre ce phénomène. Supposons que vous constituiez un échantillon à partir de la population à l'étude, dans ce cas les enfants participant au programme. Dans un premier temps, vous prélevez un échantillon de seulement deux observations. Dans ce cas, rien ne garantit que l'échantillon présente les mêmes caractéristiques que la population à l'étude. Vous risquez en effet de sélectionner deux individus présentant des caractéristiques inhabituelles. Ainsi, même si seulement 20 % des enfants de la population à l'étude portent des chapeaux ronds, il est possible que vous préleviez un échantillon de deux enfants à chapeau rond. Ce serait un coup de malchance, mais ce n'est pas impossible. Augmenter la taille de l'échantillon permet de réduire ce risque. Un grand échantillon est plus susceptible de ressembler à la population à l'étude qu'un petit échantillon. La figure 11.1 illustre ce qui se passe lorsque vous prélevez un grand échantillon. Il est fort probable qu'un grand échantillon présente plus ou moins les mêmes caractéristiques que la population : dans notre exemple, 20 % des enfants portent des chapeaux ronds, 10 % portent des chapeaux carrés et 70 % portent des chapeaux triangulaires.

Nous savons maintenant qu'un grand échantillon permet de donner une image plus précise de la population des enfants participants. Il en va de même pour les enfants non participants : plus l'échantillon de non participants est grand, plus l'image que nous obtenons de la population est précise. Pourquoi est-ce important ? Si nous sommes en mesure d'estimer le résultat (poids) moyen des enfants participants et non participants plus précisément, nous serons également en mesure d'établir avec plus de précision la différence de poids entre les deux groupes, et donc

Figure 11.1 Un grand échantillon ressemble mieux à la population



l'impact du programme. En d'autres termes, si vous n'avez qu'une vague idée du poids moyen des enfants dans les groupes de traitement et de comparaison, vous ne pourrez pas avoir une idée précise de la différence de poids entre les deux groupes. Dans la section suivante, nous développons cette idée de façon légèrement plus formelle.

Comparer les résultats moyens entre les groupes de traitement et de comparaison

Une fois que vous avez estimé le résultat (poids) moyen du groupe de traitement (enfants participants sélectionnés par assignation aléatoire) et du groupe de comparaison (enfants non participants sélectionnés par assignation aléatoire), vous pouvez déterminer s'il existe une différence entre les deux. Il vous suffit de soustraire les moyennes pour obtenir la différence. L'évaluation d'impact compare alors *l'hypothèse nulle (ou hypothèse par défaut),*

H_0 : impact = 0 (hypothèse selon laquelle le programme n'a pas d'impact),

à l'hypothèse alternative :

H_a : impact ≠ 0 (hypothèse selon laquelle le programme a un impact).

Imaginez que, dans l'exemple du programme de nutrition, vous commenciez votre évaluation à partir d'un échantillon de deux enfants traités et de deux enfants de comparaison. Au vu de la taille réduite de l'échantillon, votre estimation du poids moyen des enfants traités et des enfants constituant le groupe de comparaison, et donc de la différence entre les deux groupes, ne sera pas très fiable. Vous pouvez vérifier ce phénomène en prélevant différents échantillons de deux enfants traités et deux enfants constituant le groupe de comparaison. Vous constaterez que l'impact estimé du programme varie grandement.

Maintenant, imaginons que vous effectuiez votre évaluation à partir d'un échantillon de 1 000 enfants traités et de 1 000 enfants constituant le groupe de comparaison. Comme nous l'avons mentionné, vos estimations du poids moyen des deux groupes seront beaucoup plus précises. Votre estimation de la différence entre les deux groupes en sera également d'autant plus précise.

Admettons que vous obteniez un poids moyen pour l'échantillon des enfants traités (participants) de 25,2 kg contre 25 kg pour l'échantillon des enfants non participants (groupe de comparaison). La différence entre les deux groupes s'établit à 0,2 kg. Si ces chiffres avaient été obtenus à partir d'échantillons comprenant chacun deux observations, vous n'auriez pas pu être certain que l'impact de 0,2 kg ne provient pas d'un manque de précision dans vos estimations. En revanche, s'ils sont obtenus à partir d'échantillons de 1 000 observations chacun, vous pouvez affirmer avec plus de certitude que votre estimation se rapproche du véritable impact du programme, qui dans ce cas est positif.

La question fondamentale devient alors : *quelle taille doit précisément avoir l'échantillon pour être sûr qu'une estimation d'impact positif reflète effectivement le véritable impact du programme et non un manque de précision des estimations ?*

Deux erreurs potentielles dans les évaluations d'impact

Lorsque vous cherchez à déterminer si un programme a un impact, vous pouvez commettre deux types d'erreurs potentielles. Une *erreur de type I* apparait lorsque l'évaluation amène à conclure qu'un programme a eu un impact alors que ce n'est pas le cas. Dans notre exemple du programme de nutrition, cette erreur serait commise si, en tant qu'évaluateur, vous arriviez à la conclusion que le poids moyen des enfants de l'échantillon traité était supérieur à celui des enfants de l'échantillon de comparaison, alors que le poids moyen dans les deux groupes est en fait équivalent. Dans ce cas, l'impact positif que vous avez constaté est entièrement attribuable au manque de précision de vos estimations.

À l'inverse, une *erreur de type II* apparait lorsque l'évaluation amène à conclure qu'un programme n'a eu aucun impact alors qu'il en a en réalité eu un. Dans notre exemple du programme de nutrition, vous commettriez une erreur de type II si vous arriviez à la conclusion que le poids moyen des enfants des deux échantillons est le même, alors que le poids moyen des enfants de la population traitée est en fait différent à celui des enfants du groupe de comparaison. Là encore, l'impact estimé aurait dû être différent de zéro, mais le manque de précision de vos estimations vous amène à conclure que le programme n'a pas eu d'impact.

Concept clé :
La puissance est la probabilité d'observer un impact s'il se produit. La puissance d'une évaluation d'impact est élevée si le risque de ne pas observer un impact existant, c'est-à-dire de commettre une erreur de type II, est faible.

Lorsqu'ils testent l'hypothèse qu'un programme a eu un impact, les statisticiens peuvent limiter la probabilité d'erreurs de type I. En effet, la probabilité de commettre une erreur de type I est définie par un paramètre appelé le *« niveau de confiance »*. Le niveau de confiance est souvent fixé à 5 %, ce qui indique que vous pouvez être sûr à 95 % de votre conclusion selon laquelle le programme a eu un impact. Si vous craignez de commettre une erreur de type I, vous pouvez fixer un niveau de confiance plus faible, de 1 % par exemple, pour pouvoir être sûr à 99 % de votre conclusion selon laquelle le programme a eu un impact.

Les erreurs de type II sont également une source d'inquiétude pour les décideurs. De nombreux facteurs influencent la probabilité de commettre une erreur de type II ; toutefois, la taille de l'échantillon est un facteur déterminant. Si le poids moyen de 50 000 enfants traités est le même que le poids moyen de 50 000 enfants de comparaison, vous pouvez probablement conclure avec certitude que le programme n'a pas eu d'impact. En revanche, si les deux enfants de votre échantillon de traitement pèsent le même poids que les deux enfants du groupe de comparaison, il sera plus difficile de formuler une conclusion avec certitude. Vous vous demanderez alors si le poids moyen est similaire parce que l'intervention n'a pas eu d'impact ou parce que les données sont insuffisantes pour tester l'hypothèse à partir d'un échantillon si petit. En prélevant de grands échantillons, vous réduisez le risque de

n'observer, par (bon ou mauvais) hasard, que des enfants ayant le même poids. Avec de grands échantillons, la différence de moyennes entre un échantillon de traitement et un échantillon de comparaison fournit une estimation fiable de la véritable différence qui existe entre toutes les unités traitées et toutes les unités du groupe de comparaison.

La *puissance (ou puissance statistique)* d'une évaluation d'impact correspond à la probabilité qu'elle détecte une différence entre les groupes de traitement et de comparaison, si une telle différence existe. La puissance d'une évaluation d'impact est élevée si le risque de ne pas observer un impact qui existe, c'est-à-dire de commettre une erreur de type II, est faible. Les exemples cités plus haut montrent que la taille de l'échantillon est un facteur déterminant de la puissance d'une évaluation d'impact. Nous allons approfondir cette intuition dans les sections suivantes.

Pourquoi les calculs de puissance sont déterminants pour les décisions politiques

Les calculs de puissance permettent de déterminer la taille d'échantillon nécessaire pour éviter de conclure qu'un programme n'a pas eu d'impact alors qu'il en a en fait eu un (erreur de type II). La puissance d'un test est égale à un moins la probabilité d'une erreur de type II.

La *puissance* d'une évaluation d'impact est élevée si une erreur de type II est peu probable, c'est-à-dire qu'il y a peu de chance que vous obteniez des résultats indiquant que le programme évalué n'a pas eu d'impact alors qu'il en a bel et bien eu un.

Dans une perspective purement politique, les *évaluations d'impact de faible puissance* qui présentent un fort risque d'erreur de type II sont potentiellement non seulement inutiles, mais également très coûteuses. La forte probabilité d'erreur de type II compromet la fiabilité de tout résultat n'indiquant pas d'impact. Consacrer des ressources à des évaluations d'impact à faible puissance est donc un investissement risqué.

Les évaluations à faible puissance peuvent également avoir des conséquences dramatiques sur le plan pratique. Par exemple, dans notre exemple précédent du programme de nutrition, si vous concluez que le programme n'a pas d'impact alors qu'il en a bel et bien eu un, les décideurs seront susceptibles de mettre fin à un programme qui est en fait bénéfique pour les enfants. Il est donc essentiel de limiter la probabilité d'erreurs de type II en utilisant des échantillons assez grands dans le cadre des évaluations d'impact. C'est la raison pour laquelle il est si important et pertinent d'effectuer des calculs de puissance.

Les calculs de puissance étape par étape

Nous abordons maintenant les principes fondamentaux de calculs de puissance à partir du cas simple d'un programme assigné aléatoirement. Pour réaliser des calculs de puissance, il faut poser les six questions suivantes :

1. Le programme produit-il des *grappes* ?
2. Quel est l'indicateur de *résultat* ?

3. Souhaitez-vous comparer les impacts du programme entre plusieurs *sous-groupes* ?
4. Quel est le *niveau minimum* d'impact qui justifierait l'investissement effectué dans l'intervention ?
5. Quel est le *niveau de puissance* raisonnable pour l'évaluation réalisée ?
6. Quelles sont la *moyenne et la variance de référence des indicateurs de résultats* ?

Chacune de ces étapes doit être considérée dans le contexte politique particulier au sein duquel l'évaluation d'impact est mise en œuvre.

Nous avons déjà mentionné que l'échelle minimum d'intervention conditionne la taille de l'échantillon nécessaire pour l'évaluation. La première étape des calculs de puissance consiste à déterminer si le programme que vous voulez évaluer produit des *grappes*. Des grappes sont formées lorsque le niveau d'intervention du programme est différent du niveau auquel vous souhaitez mesurer les résultats. Par exemple, un programme peut être mis en œuvre au niveau d'un hôpital, d'une école ou d'un village (donc, par grappes) alors que vous souhaitez mesurer son impact sur les patients, les étudiants ou les villageois (voir tableau 11.1)[6].

La nature des données d'échantillons obtenues pour un programme formant des grappes diffère légèrement de celle des échantillons obtenus pour un programme qui ne forme pas de grappes. Par conséquent, les étapes des calculs de puissance sont légèrement différentes selon que le traitement est assigné de manière aléatoire aux différentes grappes ou à toutes les unités d'une population. Nous aborderons chaque situation l'une après l'autre. Nous commencerons par les principes des calculs de puissance en l'absence de grappes, c'est-à-dire lorsque le traitement est assigné au niveau où les résultats sont observés, avant de passer aux cas où il existe des grappes.

Tableau 11.1 Exemples de grappes

Traitement	Niveau d'assignation du traitement (grappe)	Unité auprès de laquelle le résultat est mesuré
Transferts monétaires conditionnels	Village	Ménages
Traitement contre le paludisme	École	Individus
Programme de formation	Quartier	Individus

Calculs de puissance en absence de grappes

Admettons que vous ayez répondu à la première question ci-dessus en déterminant que les bénéfices du programme à évaluer ne sont pas assignés par grappe. En d'autres termes, les bénéfices du programme sont assignés de manière aléatoire à toutes les unités de la population éligible. Dans ce cas, l'échantillon d'évaluation peut être constitué en prélevant un échantillon aléatoire de la *population à l'étude*.

Les deuxième et troisième étapes concernent les objectifs de l'évaluation. À la deuxième étape, vous devez déterminer les principaux *indicateurs de résultats* que le programme cherche à améliorer. Ces indicateurs découlent de la question de recherche fondamentale de l'évaluation et du cadre conceptuel décrits dans la Partie 1. La présente discussion contribuera à illustrer les types d'indicateurs qui se prêtent le mieux à une évaluation d'impact.

Troisièmement, la question de politique qui motive l'évaluation requiert parfois la comparaison des impacts d'un programme entre différents *sous-groupes*, par exemple entre individus de différents âges ou catégories de revenus. Si tel est le cas, la taille des échantillons devra être plus grande, et les calculs de puissance devront être adaptés en conséquence. Par exemple, la question de politique peut demander si l'impact d'un programme éducatif varie entre les filles et les garçons. Intuitivement, il faut disposer d'un nombre suffisant d'étudiants de chaque genre au sein du groupe de traitement et du groupe de comparaison pour pouvoir détecter l'impact sur chaque sous-groupe. Pour comparer l'impact d'un programme entre deux sous-groupes, il est parfois nécessaire de doubler la taille de l'échantillon. Lorsque les sous-groupes sont plus hétérogènes (tranches d'âges par exemple), la taille de l'échantillon nécessaire peut considérablement augmenter.

Quatrièmement, vous devez déterminer l'impact minimum qui justifierait l'investissement effectué dans l'intervention. Il s'agit fondamentalement d'une question de politique et non d'une question technique. Un programme de transferts monétaires conditionnels est-il un investissement justifié s'il réduit la pauvreté de 5 %, de 10 % ou de 15 % ? Un programme d'activation sur le marché du travail se justifie-t-il s'il augmente les revenus de 5 %, de 10 % ou de 15 % ? La réponse dépend du contexte, mais il est nécessaire, quelles que soient les circonstances, de déterminer le changement des indicateurs de résultat qui justifierait un investissement dans le programme. Autrement dit, *il faut déterminer le niveau d'impact en dessous duquel une intervention est considérée comme un échec ?* La réponse à cette question dépend non seulement du coût du programme et du type de traitement qu'il propose, mais également du coût d'opportunité de ne pas investir les fonds dans une autre intervention.

Les calculs de puissance permettent d'adapter la taille de l'échantillon pour pouvoir détecter l'effet minimal désiré. Pour qu'une évaluation détecte un faible impact, les estimations de la différence de résultats moyens entre le groupe de traitement et le groupe de comparaison devront être très précises, ce qui nécessite un grand échantillon. En revanche, pour les interventions qui ne sont jugées utiles que si elles entraînent des changements considérables des indicateurs de résultat, les échan-

tillons utilisés pour l'évaluation d'impact pourront être plus petits. Il convient néanmoins de déterminer *l'effet minimal détectable* avec prudence étant donné que tout impact inférieur à l'effet minimal désiré risque de ne pas être détecté.

Cinquièmement, l'évaluateur doit consulter des statisticiens pour déterminer un *niveau de puissance* raisonnable pour l'évaluation prévue. Comme nous l'avons mentionné, la puissance d'un test est égale à un moins la probabilité d'une erreur de type II. La puissance est donc comprise entre zéro et un et plus elle est élevée, moins il y a de risque de ne pas détecter un impact existant. Une puissance de 80 % est un niveau fréquemment utilisé pour les calculs de puissance. Cela signifie que vous allez détecter un impact existant dans 80 % des cas. Un niveau de puissance plus élevé de 0,9 (ou 90 %) constitue souvent un niveau utile, bien que plus prudent et entraînant par conséquent une hausse de la taille de l'échantillon requis[7].

Sixièmement, vous devez demander à un statisticien d'estimer certains paramètres de référence comme la *moyenne et la variance des indicateurs de résultats*. Ces valeurs de référence doivent de préférence être obtenues à partir de données existantes collectées dans un contexte comparable à celui du programme à l'étude[8]. Il est très important de noter que plus un indicateur de résultat est variable, plus il sera difficile de formuler une estimation fiable de l'effet du traitement. Dans l'exemple du programme de nutrition, le poids des enfants est le résultat à l'étude. Si tous les enfants pèsent le même poids dans l'enquête de référence, il sera possible d'estimer l'impact de l'intervention à partir d'un échantillon relativement réduit. En revanche, si les poids de référence des enfants affichent une grande variance, un échantillon plus grand sera nécessaire pour estimer l'impact du programme.

À l'issue de ces six étapes, le statisticien peut effectuer le calcul de puissance en utilisant un logiciel statistique standard[9]. Le calcul de puissance qui en résultera indiquera la taille de l'échantillon nécessaire en fonction des paramètres définis aux étapes 1 à 6. Les calculs de puissance sont faciles à réaliser une fois que les questions d'ordre politique (points 3 et 4) ont été résolues.[10]

Lorsqu'un statisticien est mandaté pour faire des calculs de puissance, il est recommandé de demander une analyse de la *sensibilité* des calculs de puissance aux changements d'hypothèses. Ceci est important pour comprendre dans quelle mesure la taille de l'échantillon devra être augmentée pour que les hypothèses de départ deviennent plus conservatrices (baisse de l'impact espéré, hausse de la variance dans l'indicateur de résultat ou niveau de puissance plus élevé). Il est également utile de demander des calculs de puissance pour différents indicateurs de résultats étant donné que la taille de l'échantillon nécessaire peut considérablement changer si certains indicateurs de résultats sont plus ou moins variables que d'autres.

Enfin, les calculs de puissance permettent d'établir la taille minimum de l'échantillon nécessaire. Dans la pratique, les problèmes de mise en œuvre impliquent souvent que la taille de l'échantillon effectif soit inférieure à la taille prévue. Toute déviation de ce type doit être envisagée avec prudence, mais il est conseillé d'ajouter une marge de 10 à 20 % à la taille de l'échantillon prévue par les calculs de puissance[11].

Concept clé :

La taille de l'échantillon requis augmente si l'effet minimal détectable est faible, si l'indicateur de résultat est très variable ou s'il s'agit d'un événement rare, et si l'évaluation vise à comparer les impacts entre différents sous-groupes.

Taille de l'échantillon nécessaire pour évaluer une version amplifiée du Programme de subvention de l'assurance maladie (PSAM)

La présidente et le ministre de la Santé se sont montrés satisfaits de la qualité et des résultats de l'évaluation du Programme de subvention de l'assurance maladie (PSAM), notre exemple des chapitres précédents. Toutefois, avant d'étendre le PSAM, ils décident de mettre en œuvre à titre de projet pilote une version amplifiée du programme, qu'ils appellent PSAM+. Le PSAM finance une partie des frais de l'assurance maladie pour les ménages pauvres qui vivent en milieu rural, couvrant les dépenses relatives aux soins de santé primaires et à l'achat de médicaments, mais pas les frais d'hospitalisation. La présidente et le ministre de la Santé souhaitent savoir si un PSAM+ amplifié couvrant également les frais d'hospitalisation pourrait permettre de réduire davantage les dépenses de santé à la charge directe des ménages. Ils vous demandent donc de concevoir une évaluation d'impact pour savoir si le PSAM+ réduit bel et bien les dépenses de santé des ménages ruraux pauvres.

Dans ce contexte, vous n'hésitez pas sur le choix de la méthode d'évaluation d'impact : le PSAM+ est doté de ressources limitées. Dans l'immédiat, il ne peut pas être mis en œuvre auprès de l'ensemble de la population. Vous concluez donc que l'assignation aléatoire est la méthode d'évaluation la plus pertinente et la plus robuste. La présidente et le ministre de la Santé comprennent le fonctionnement de la méthode d'assignation aléatoire et y sont très favorables.

Afin de finaliser la conception de l'évaluation d'impact, vous demandez à un statisticien de vous aider à définir la taille de l'échantillon nécessaire. Avant de commencer, le statisticien vous demande de lui fournir certaines informations clés. Il a six questions à vous poser.

1. Le statisticien demande si le programme PSAM+ va générer des grappes. À ce stade, vous ne le savez pas encore. Vous pensez qu'il est possible de procéder à une assignation aléatoire du PSAM + au niveau des ménages parmi tous les ménages ruraux pauvres qui bénéficient déjà du PSAM. Toutefois, il vous semble possible que la présidente et le ministre de la Santé préfèrent peut-être assigner le programme au niveau des villages, ce qui entraînerait la création de grappes. Le statisticien propose de commencer par effectuer des calculs de puissance sans grappe, puis d'examiner dans quelle mesure l'existence de grappes influerait sur les résultats.

2. Le statisticien vous demande quel est l'indicateur de résultat. Vous expliquez que le gouvernement souhaite utiliser un indicateur bien défini : les dépenses de santé directes des ménages. Le statisticien cherche une base de données récente pour obtenir des valeurs de référence pour cet indicateur. Il propose d'utiliser l'enquête de suivi de l'évaluation du PSAM. Il remarque que, parmi les ménages ayant bénéficié du PSAM, les dépenses annuelles de santé directes par personne s'élèvent en moyenne à 7,84 dollars.

3. Le statisticien s'assure que vous ne souhaitez pas mesurer l'impact du programme sur des sous-groupes par exemple des régions ou des populations spécifiques.

4. Le statisticien demande quel est l'impact minimum qui justifierait un investissement dans la version amplifiée du programme. En d'autres termes, il veut connaître le montant de la baisse des dépenses de santé en dessous de la moyenne de référence de 7,84 dollars qui justifierait l'intervention. Il explique qu'il ne s'agit pas selon lui d'une considération technique, mais plutôt d'une question d'ordre politique. Pour cette raison c'est à un décideur tel que vous de déterminer l'effet minimum que l'évaluation doit permettre de détecter. Vous avez entendu la présidente mentionner que le PSAM+ serait considéré comme efficace s'il permettait de réduire les dépenses de santé directes des ménages de deux dollars. Toutefois, vous savez que, dans le cadre de l'évaluation, il vaut mieux être prudent dans la détermination de l'impact minimum détectable, tout impact inférieur étant peu susceptible d'être détecté. Pour comprendre comment la taille de l'échantillon nécessaire varie en fonction de l'effet minimum détectable, vous suggérez au statisticien d'effectuer des calculs en vue d'une réduction minimum des dépenses de santé directes de un dollar, de deux dollars et de trois dollars.

5. Le statisticien vous demande le niveau de puissance que vous jugeriez raisonnable pour l'évaluation réalisée. Il ajoute que les calculs de puissance sont généralement réalisés sur la base d'une puissance de 0,9, mais il propose de réaliser ultérieurement des tests de sensibilité à un niveau moins conservateur de 0,8.

6. Enfin, le statisticien demande quelle est la variance de l'indicateur de résultat dans la population à l'étude. Il consulte à nouveau les données des ménages ayant bénéficié du PSAM en indiquant que l'écart-type des dépenses de santé directes est de huit dollars.

Avec toutes ces informations, le statisticien effectue les calculs de puissance. Comme convenu, il commence par le cas le plus conservateur d'une puissance de 0,9. Il obtient les résultats figurant dans le tableau 11.2.

Il conclut que pour détecter une baisse de deux dollars des dépenses de santé directes avec une puissance de 0,9, l'échantillon doit contenir au moins 672 unités (336 unités traitées et 336 unités de comparaison, en l'absence de grappes). Il indique que s'il vous convenait de détecter une baisse de trois dollars des dépenses de santé directes, un échantillon plus réduit d'au moins 300 unités (150 unités dans chaque groupe) serait suffisant. En revanche, un échantillon beaucoup plus important d'au moins 2 688 unités (1 344 dans chaque groupe) serait nécessaire pour détecter une baisse de un dollar dans les dépenses de santé directes.

Tableau 11.2 Taille de l'échantillon nécessaire selon les différents effets minimums détectables (baisse des dépenses de santé des ménages), puissance = 0,9, sans grappe

Effet minimal détectable	Groupe de traitement	Groupe de comparaison	Échantillon total
$1	1 344	1 344	2 688
$2	336	336	672
$3	150	150	300

Remarque : l'effet minimal détectable correspond à la réduction minimum des dépenses de santé directes des ménages que l'évaluation d'impact doit pouvoir détecter.

Le statisticien produit ensuite un deuxième tableau pour un niveau de puissance de 0,8. Le tableau 11.3 montre que les tailles d'échantillons nécessaires sont inférieures pour une puissance de 0,8 que pour une puissance de 0,9. Pour détecter une baisse de deux dollars des dépenses de santé directes des ménages, un échantillon total d'au moins 502 unités est suffisant. Pour détecter une baisse de trois dollars, au moins 224 unités sont nécessaires. Toutefois, pour détecter une baisse de un dollar, au moins 2 008 unités sont nécessaires.

Le statisticien explique que les résultats sont typiques des calculs de puissance :

- Plus le niveau de puissance est élevé (ou prudent), plus la taille de l'échantillon nécessaire est importante.
- Plus l'impact à détecter est réduit, plus l'échantillon nécessaire est grand.

Tableau 11.3 Taille de l'échantillon nécessaire selon les différents effets minimums détectables (baisse des dépenses de santé des ménages), puissance = 0,8, sans grappe

Effet minimal détectable	Groupe de traitement	Groupe de comparaison	Échantillon total
1 $	1 004	1 004	2 008
2 $	251	251	502
3 $	112	112	224

Remarque : l'effet minimal détectable correspond à la réduction minimum des dépenses de santé directes des ménages que l'évaluation d'impact doit pouvoir détecter.

Tableau 11.4 Taille de l'échantillon nécessaire pour détecter différents effets minimums désirés (hausse du taux d'hospitalisation), puissance = 0,9, sans grappe

Effet minimal détectable (point de pourcentage)	Groupe de traitement	Groupe de comparaison	Échantillon total
1	9 717	9 717	19 434
2	2 430	2 430	4 860
3	1 080	1 080	2 160

Remarque : l'effet minimal désiré correspond au changement minimum du taux d'hospitalisation (exprimé en points de pourcentage) que l'évaluation d'impact doit pouvoir détecter.

Le statisticien demande si vous souhaitez réaliser des calculs de puissance pour d'autres indicateurs de résultats. Vous suggérez d'évaluer également la taille de l'échantillon nécessaire pour détecter si le PSAM+ affecte le taux d'hospitalisation. Dans l'échantillon des villages bénéficiant du PSAM, 5 % des ménages comptent un membre qui a été hospitalisé au cours de la dernière année. Le statisticien produit un nouveau tableau qui indique que des échantillons relativement importants seraient nécessaires même pour détecter de grands changements du taux d'hospitalisation de un, deux ou trois points par rapport au taux de référence de 5 % (tableau 11.4).

Le tableau indique que les tailles des échantillons nécessaires sont plus importantes pour ce résultat (taux d'hospitalisation) que pour les dépenses de santé directes. Le statisticien conclut que si vous souhaitez détecter les impacts sur ces deux indicateurs de résultats, vous devrez utiliser les plus grands échantillons proposés par les calculs de puissance portant sur le taux d'hospitalisation. Si vous choisissez d'utiliser des échantillons de la taille suggérée par les calculs de puissance effectués pour les dépenses de santé directes, le statisticien recommande de préciser à la présidente et au ministre de la Santé que l'évaluation ne présentera pas une puissance suffisante pour détecter les effets sur le taux d'hospitalisation.

QUESTION 8

A. Quelle taille d'échantillon recommanderiez-vous pour estimer l'impact du PSAM+ sur les dépenses de santé directes ?

B. La taille de cet échantillon est-elle suffisante pour détecter un changement du taux d'hospitalisation ?

Calculs de puissance avec grappes

Les paragraphes précédents présentent les calculs de puissance pour des programmes ne produisant pas de grappes. Toutefois, comme nous l'avons vu dans la deuxième partie du manuel, les bénéfices de nombreux programmes sont assignés par grappes. Nous allons donc décrire brièvement comment adapter les principes de base des calculs de puissance aux échantillons par grappes.

Lorsqu'il existe des grappes, il convient de noter que le nombre de grappes est un paramètre beaucoup plus important que le nombre d'individus qui composent les grappes. Un nombre suffisant de grappes est nécessaire pour pouvoir identifier avec un degré de certitude suffisant l'éventuel impact d'un programme en comparant les résultats des échantillons de traitement et de comparaison.

Si vous assignez de façon aléatoire un traitement au sein d'un petit nombre de grappes, il est peu probable que les groupes de traitement et de comparaison soient identiques. L'assignation aléatoire entre deux districts, deux écoles ou deux hôpitaux ne garantit pas que les deux grappes soient similaires. En revanche, l'assignation aléatoire d'une intervention entre 100 districts, 100 écoles ou 100 hôpitaux a plus de probabilités de créer un groupe de traitement et un groupe de comparaison similaires. En résumé, un nombre suffisant de grappes est nécessaire pour s'assurer qu'un équilibre est atteint. Par ailleurs, le nombre de grappes joue également un rôle dans la précision des effets estimés. Un nombre suffisant de grappes est nécessaire pour tester l'hypothèse selon laquelle un programme a un impact avec une puissance satisfaisante. Il est donc très important de s'assurer que le nombre de grappes disponibles pour l'assignation aléatoire est assez grand.

Sur la base de l'intuition décrite ci-dessus, vous pouvez définir le nombre de grappes nécessaire pour effectuer un test d'hypothèse précis en effectuant des calculs de puissance. La réalisation de calculs de puissance pour des échantillons à grappes nécessite une étape supplémentaire par rapport à la procédure de base :

1. Le programme produit-il des grappes ?
2. Quel est l'indicateur de résultat ?
3. Souhaitez-vous comparer les impacts du programme entre plusieurs sous-groupes ?
4. Quel est le niveau minimum d'impact qui justifierait l'investissement effectué dans l'intervention ?
5. Quelle est la moyenne de référence de l'indicateur de résultat ?
6. Quelle est la variance de l'indicateur de résultat dans la population à l'étude ?
7. Quelle est la variance de l'indicateur de résultat au sein des grappes ?

Par rapport aux calculs de puissance sans grappe, une étape est ajoutée : vous devez demander à votre statisticien quel est le degré de corrélation entre les résultats au sein des grappes. À l'extrême, tous les résultats d'une même grappe peuvent afficher une corrélation parfaite. Par exemple, il est possible que les revenus des ménages ne soient pas particulièrement variables au sein d'un village, mais que d'importantes inégalités existent entre différents villages. Dans ce cas, si vous souhaitez ajouter un individu à votre échantillon d'évaluation, l'addition d'un individu d'un nouveau village augmentera plus la puissance que l'addition d'un individu

venant d'un village déjà représenté. En effet, dans ce dernier cas, le deuxième villageois sera très probablement similaire au villageois déjà inclus dans l'échantillon. En général, *plus la corrélation intra-grappe* des résultats est élevée, plus le nombre de grappes nécessaire pour obtenir un niveau de puissance donné augmente.

Dans les échantillons par grappes, les calculs de puissance mettent en évidence la balance nécessaire entre l'ajout de nouvelles grappes à l'échantillon et de nouvelles observations dans les grappes de l'échantillon. L'augmentation relative de la puissance due à l'ajout d'une unité au sein d'une nouvelle grappe est presque toujours plus importante que l'ajout d'une unité au sein d'une grappe existante. Bien que l'augmentation de puissance associée à l'ajout d'une nouvelle grappe puisse être importante, l'ajout de grappes peut aussi avoir des conséquences opérationnelles et affecter le coût de la collecte de données. La section suivante explique comment réaliser des calculs de puissance avec des grappes dans l'exemple du PSAM+ et certaines des décisions possibles.

Concept clé :
Pour les calculs de puissance, le nombre de grappes est plus important que le nombre d'individus au sein des grappes. Il faut le plus souvent au moins 30 grappes dans chaque groupe (groupe de traitement et groupe de comparaison).

Dans de nombreux cas, il faut au moins 30 à 50 grappes dans le groupe de traitement et dans le groupe de comparaison pour obtenir une puissance suffisante et garantir l'équilibre des caractéristiques de référence avec la méthode d'assignation aléatoire. Toutefois, le nombre peut varier en fonction des différents paramètres mentionnés ci-dessus, tout comme le degré de corrélation intra-grappe. De plus, le nombre de grappes nécessaire est généralement plus élevé avec des méthodes autres que l'assignation aléatoire (toutes choses égales par ailleurs).

Taille de l'échantillon nécessaire pour évaluer une version amplifiée du Programme de subvention de l'assurance maladie (PSAM) avec grappes

Après votre discussion avec le statisticien concernant les calculs de puissance pour le PSAM+, vous décidez de vous entretenir avec la présidente et le ministre de la Santé au sujet des conséquences d'une assignation aléatoire du PSAM+ aux individus bénéficiant déjà du PSAM. Cette conversation vous permet d'établir qu'une telle procédure ne serait pas politiquement réalisable : il serait difficile d'expliquer pourquoi une personne pourrait bénéficier d'une couverture supplémentaire, mais pas son voisin.

Au lieu d'appliquer la sélection aléatoire au niveau individuel, vous proposez de sélectionner de manière aléatoire plusieurs villages bénéficiant du PSAM pour piloter le PSAM+. Tous les habitants des villages sélectionnés seraient alors éligibles. Cette procédure entraîne la création de grappes et nécessite donc de nouveaux calculs de puissance. Vous cherchez maintenant à déterminer la taille de l'échantillon nécessaire pour évaluer l'impact du PSAM+ dans un contexte d'assignation aléatoire par grappe.

Vous consultez à nouveau votre statisticien. Il vous rassure en déclarant que cela ne nécessite qu'un petit effort supplémentaire. Sur sa liste, seule une question reste sans réponse. Il doit savoir la variance de l'indicateur de résultat au sein des grappes. Il trouve la réponse à cette question dans les données de suivi du PSAM : la corrélation intra-village des dépenses de santé directes est égale à 0,04.

Il vous demande également si un plafond a été défini pour le nombre de villages dans lesquels le nouveau projet pilote peut être lancé. Étant donné que le PSAM a été mis en œuvre dans 100 villages, vous lui expliquez que vous pourriez avoir au maximum 50 villages de traitement et 50 villages de comparaison pour le programme PSAM+. Sur la base de ces informations, le statisticien produit les calculs illustrés au tableau 11.5 pour une puissance de 0,9.

Il conclut que pour détecter une baisse de deux dollars des dépenses de santé directes, l'échantillon doit inclure au moins 900 unités, soit neuf unités par grappe pour 100 grappes. Il note que ce chiffre est supérieur à celui de l'échantillon correspondant à l'assignation aléatoire au niveau des ménages qui ne nécessite que 672 unités au total. Pour détecter une baisse de trois dollars des dépenses de santé directes, l'échantillon doit inclure au moins 340 unités, ou quatre unités par grappe pour 85 grappes.

Toutefois, lorsque le statisticien tente de définir l'échantillon nécessaire pour détecter une baisse de un dollar des dépenses de santé directes, il conclut qu'un tel impact ne pourrait pas être détecté avec 100 grappes. Au moins 109 grappes seraient nécessaires, et le nombre d'observations au sein de chaque grappe serait extrêmement élevé. Ces conclusions indiquent qu'un grand nombre de grappes est nécessaire pour qu'une évaluation ait assez de puissance pour détecter des impacts relativement réduits, indépendamment du nombre d'observations au sein de chaque grappe.

Le statisticien suggère alors de recalculer ces chiffres avec une puissance de seulement 0,8 (voir le tableau 11.6). Les tailles d'échantillon nécessaires sont plus réduites, mais restent plus importantes pour un échantillon par grappes que pour un simple échantillon aléatoire.

Tableau 11.5 Taille de l'échantillon nécessaire pour différents effets minimums détectables (baisse des dépenses de santé des ménages), puissance = 0,9, 100 grappes maximum

Effet minimal détectable	Nombre de grappes	Unités par grappe	Échantillon total avec grappes	Échantillon total sans grappe
1 $	Impossible	Impossible	Impossible	2 688
2 $	100	9	900	672
3 $	85	4	340	300

Remarque : l'effet minimal détectable correspond à la réduction minimum des dépenses de santé directes des ménages que l'évaluation d'impact doit pouvoir détecter.

Tableau 11.6 Taille de l'échantillon nécessaire pour différents effets minimums détectables (baisse des dépenses de santé des ménages), puissance = 0,8, 100 grappes maximum

Effet minimal détectable	Nombre de grappes	Unités par grappe	Échantillon total avec grappes	Échantillon total sans grappe
$1	100	102	10 200	2 008
$2	90	7	630	502
$3	82	3	246	224

Remarque : l'effet minimal détectable correspond à la réduction minimum des dépenses de santé directes des ménages que l'évaluation d'impact doit pouvoir détecter.

Le statisticien calcule alors comment le nombre total d'observations nécessaires varie en fonction du nombre de grappes. Il décide de refaire les calculs pour un effet minimal détectable de deux dollars et une puissance de 0,9. La taille de l'échantillon total nécessaire pour estimer un tel impact augmente fortement lorsque le nombre de grappes diminue (tableau 11.7). Pour 100 grappes, 900 observations sont nécessaires. Pour 30 grappes, l'échantillon total doit alors contenir 6 690 observations. En revanche, pour 157 grappes, seules 785 observations sont nécessaires.

QUESTION 9

A. Quelle taille d'échantillon total recommanderiez-vous pour estimer l'impact du PSAM+ sur les dépenses de santé directes ?

B. Dans combien de villages recommanderiez-vous à la présidente et au ministre de la Santé de déployer le PSAM+ ?

Tableau 11.7 Taille de l'échantillon nécessaire pour détecter un impact minimum de 2 dollars pour différents nombres de grappes, puissance = 0,9

Effet minimal détectable	Nombre de grappes	Unités par grappe	Échantillon total sans grappe
2 $	30	223	6 690
2 $	60	20	1 200
2 $	86	11	946
2 $	100	9	900
2 $	120	7	840
2 $	135	6	810
2 $	157	5	785

En résumé

Pour résumer, la qualité d'une évaluation d'impact dépend directement de la qualité des données sur lesquelles elle se fonde. À ce titre, il est essentiel de créer des échantillons bien construits et d'une taille adéquate. Nous avons passé en revue les principes de base des calculs de puissance. Dans la planification d'une évaluation, les calculs de puissance sont un outil essentiel pour limiter les coûts de collecte des données. Ils permettent d'éviter de collecter plus de données que nécessaire tout en minimisant le risque de conclure de façon erronée qu'un programme n'a pas d'impact alors qu'il en a bien eu un. Les calculs de puissance se fondent sur des informations techniques et statistiques, mais aussi sur des décisions politiques. En général, l'augmentation de la taille de l'échantillon a des rendements décroissants. Pour définir la taille de l'échantillon adéquate, il faut donc trouver le juste équilibre entre la précision des estimations d'impact et les considérations budgétaires.

Concept clé :
Les méthodes d'évaluation d'impact quasi expérimentales nécessitent presque toujours des échantillons plus grands que le cas de référence de l'assignation aléatoire.

Nous nous sommes concentrés ici sur le cas de référence d'une évaluation d'impact mise en œuvre au moyen de l'assignation aléatoire. Il s'agit du scénario le plus simple, et donc le plus adapté pour décrire l'intuition sous-jacente aux calculs de puissance. De nombreux aspects pratiques des calculs de puissance n'ont cependant pas été abordés, et de nombreux scénarios divergent des exemples simplifiés présentés ici. Par exemple, les méthodes d'évaluation d'impact quasi expérimentales nécessitent presque toujours des échantillons plus importants que dans le cas de l'assignation aléatoire. Par ailleurs, la taille de l'échantillon nécessaire augmente s'il existe un risque de biais dans l'estimation des effets du traitement ou dans les cas où l'adhérence n'est pas totale. Ces aspects sortent du cadre du présent ouvrage, mais vous trouverez une description plus détaillée dans Spybrook et al. (2008) ou Rosenbaum (2009, chapitre 14). Il existe plusieurs ressources de référence pour approfondir la conception d'échantillons. Ainsi, la Fondation W.T. Grant a mis au point l'*Optimal Design Software for Multi-Level and Longitudinal Research*, un logiciel utile pour les analyses de puissance statistique en présence de grappes. Dans la pratique, les agences qui commanditent une évaluation sont nombreuses à faire appel à un spécialiste pour effectuer les calculs de puissance. Ce dernier devrait être à même de prodiguer des conseils si des méthodes autres que l'assignation aléatoire sont utilisées.

Choisir une stratégie d'échantillonnage

La taille n'est pas le seul facteur qui garantit l'adéquation d'un échantillon pour une évaluation d'impact. Le procédé utilisé pour prélever l'échantillon de la population à l'étude a également une grande importance. Les principes d'échantillonnage peuvent orienter le prélèvement d'échantillons représentatifs. L'échantillonnage comprend trois étapes :

1. Déterminer la *population à l'étude*.
2. Définir un *cadre d'échantillonnage*.

3. *Prélever* autant d'unités du cadre d'échantillonnage que les calculs de puissance le nécessitent.

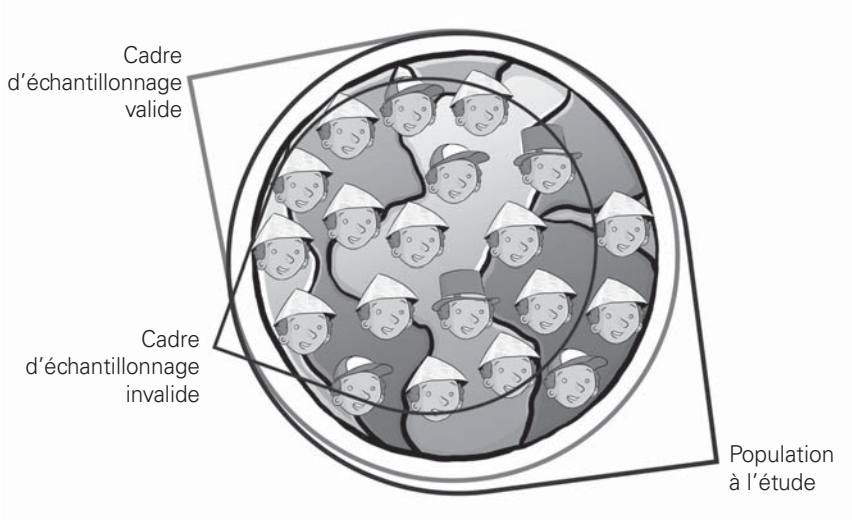
Dans un premier temps, la *population à l'étude* doit être clairement définie[12]. Pour cela, il convient de définir avec précision l'unité d'observation pour laquelle les résultats seront mesurés, avec une description claire de la couverture géographique ou de tout autre attribut pertinent caractérisant la population. Par exemple, si vous gérez un programme de développement de la petite enfance, vous pouvez chercher à mesurer les résultats cognitifs pour des enfants de trois à six ans dans l'ensemble du pays, pour des enfants de cette tranche d'âges uniquement dans les zones rurales ou seulement pour des enfants inscrits à l'école maternelle.

Dans un deuxième temps, une fois que la population à l'étude a été définie, il convient de créer un *cadre d'échantillonnage*. Le cadre d'échantillonnage est la liste la plus exhaustive qui puisse être dressée des unités d'une population à l'étude. Théoriquement, le cadre d'échantillonnage doit coïncider parfaitement avec la population à l'étude. Par exemple, un recensement parfaitement à jour de la population à l'étude constituerait un cadre d'échantillonnage idéal. Dans la pratique, des listes existantes comme les données d'un recensement de population, d'un recensement institutionnel ou des listes d'inscription à un programme sont souvent utilisées comme cadres d'échantillonnage.

Concept clé :
Le cadre d'échantillonnage est la liste existante la plus exhaustive des unités constituant la population à l'étude. Un biais de couverture apparait s'il y a une divergence entre le cadre d'échantillonnage et la population à l'étude.

Un bon cadre d'échantillonnage est essentiel pour que les conclusions tirées de l'analyse d'un échantillon soient applicables à l'ensemble de la population. En effet, un cadre d'échantillonnage qui ne coïncide pas parfaitement avec la population à l'étude engendre un *biais de couverture*, comme l'illustre la figure 11.2. En présence d'un biais de couverture, les résultats de l'échantillon n'ont pas une validité externe pour l'ensemble de la population à l'étude, mais uniquement pour

Figure 11.2 Un cadre d'échantillonnage valide couvre l'intégralité de la population à l'étude



la population du cadre d'échantillonnage. Par conséquent, les biais de couverture faussent l'interprétation des résultats de l'évaluation d'impact puisque la source de ces résultats n'est pas claire.

Lorsque vous envisagez de prélever un nouvel échantillon ou d'évaluer la qualité d'un échantillon existant, il est important de déterminer si le meilleur cadre d'échantillonnage disponible coïncide avec la population à l'étude. La généralisation des statistiques extraites de l'échantillon à toute la population à l'étude dépend de l'ampleur du biais de couverture, autrement dit de l'absence de différence entre le cadre d'échantillonnage et la population à l'étude.

Par exemple, un biais de couverture peut apparaitre si vous souhaitez étudier tous les ménages d'un pays, mais que vous utilisez l'annuaire téléphonique comme cadre d'échantillonnage : dans ce cas, les ménages sans téléphone ne seront pas inclus dans l'échantillon. Ceci peut fausser les résultats de l'évaluation si les ménages sans téléphone présentent également d'autres caractéristiques qui les différencient de la population à l'étude et que ces caractéristiques affectent la façon dont les ménages bénéficieraient de l'intervention. Par exemple, les ménages sans téléphone peuvent se situer dans des zones rurales reculées. Si vous souhaitez évaluer l'impact d'un programme de formation professionnelle, omettre les ménages les plus isolés peut affecter les résultats de l'évaluation, car ces ménages sont probablement ceux qui ont le plus de mal à intégrer le marché du travail.

Le risque de biais de couverture est réel, et la prudence est donc de rigueur lors de la définition des cadres d'échantillonnage. Par exemple, les données de recensement peuvent contenir la liste de toutes les unités d'une population. Toutefois, si une période trop longue s'est écoulée entre le recensement et la formation d'un échantillon, le cadre d'échantillonnage peut ne pas être totalement à jour, ce qui créera un biais de couverture. Par ailleurs, il est possible que les données de recensement ne contiennent pas suffisamment d'informations sur des caractéristiques précises pour pouvoir constituer un cadre d'échantillonnage. Si la population à l'étude est constituée d'enfants allant à l'école maternelle et que le recensement ne contient pas d'informations sur les inscriptions à l'école, des données complémentaires seront nécessaires[13].

Une fois que la population à l'étude et le cadre d'échantillonnage sont définis, vous devez choisir la méthode de prélèvement de l'échantillon. Il existe plusieurs procédures. Les méthodes d'*échantillonnage probabilistes* sont les plus rigoureuses, car elles attribuent à chaque unité une probabilité bien définie d'être sélectionnée. Les trois principales méthodes d'échantillonnage probabilistes sont les suivantes[14] :

Concept clé :
L'échantillonnage est le processus par lequel les unités sont prélevées du cadre d'échantillonnage. L'échantillonnage probabiliste attribue à chaque unité une probabilité bien définie d'être sélectionnée.

- *Échantillonnage aléatoire.* Toutes les unités de la population ont exactement la même probabilité d'être prélevées[15].
- *Échantillonnage aléatoire stratifié.* La population est divisée en groupes (hommes et femmes par exemple) et un échantillonnage aléatoire est effectué au sein de chaque groupe. Par conséquent, toutes les unités d'un même groupe (ou strate) ont la même probabilité d'être prélevées. Si les groupes sont assez grands, l'échantillonnage stratifié permet de tirer des conclusions sur les résultats non seulement au niveau de la population, mais également au sein de chaque groupe.

La stratification est essentielle pour les évaluations qui visent à comparer les impacts d'un programme entre différents sous-groupes.

- *Échantillonnage par grappes.* Les unités sont divisées en grappes et un échantillon aléatoire de grappes est prélevé. L'ensemble des unités des grappes prélevées constitue alors l'échantillon ou seul un certain nombre d'unités sont sélectionnées de manière aléatoire au sein de chaque grappe. Par conséquent, chaque grappe a une probabilité bien définie d'être sélectionnée, et les unités sélectionnées de chaque grappe ont elles aussi une probabilité bien définie d'être prélevées.

Dans le contexte d'une évaluation d'impact, la procédure de prélèvement d'un échantillon dépend souvent des règles d'éligibilité du programme à évaluer. Comme nous l'avons mentionné dans la section consacrée à la taille des échantillons, si la plus petite unité de mise en œuvre viable est plus grande que l'unité d'observation, l'assignation aléatoire du traitement engendrera la création de grappes. Pour cette raison, l'échantillonnage par grappes est souvent utilisé dans les études d'évaluation d'impact.

L'échantillonnage non probabiliste peut entraîner de graves erreurs d'échantillonnage. L'*échantillonnage dirigé* ou l'*échantillonnage de commodité* sont parfois utilisés à la place des procédures d'échantillonnage probabilistes décrites ci-dessus. Dans ces cas, des erreurs d'échantillonnage peuvent survenir même si le cadre d'échantillonnage couvre l'ensemble de la population et qu'il n'existe aucun biais de couverture. Considérons par exemple que pour une enquête nationale, un groupe d'enquêteurs est mandaté de collecter des données sur les ménages en se rendant dans les foyers les plus proches de l'école dans chaque village. En suivant cette procédure d'échantillonnage non probabiliste, il est probable que l'échantillon ne sera pas représentatif de l'ensemble de la population à l'étude. Un biais de couverture sera créé, car les foyers éloignés ne seront pas couverts par l'enquête.

En fin de compte, il faut choisir avec prudence son cadre d'échantillonnage et sa procédure d'échantillonnage pour assurer la validité externe des résultats obtenus pour l'ensemble de la population à l'étude. Même si le cadre d'échantillonnage présente une couverture parfaite et qu'une procédure d'échantillonnage probabiliste est utilisée, des erreurs non liées à l'échantillonnage peuvent affecter la validité externe de l'échantillon. Nous abordons ces erreurs dans le prochain chapitre.

Notes

1. Les données sur les coûts sont également nécessaires pour l'analyse coût-bénéfice.
2. Pour une description détaillée des enquêtes auprès des ménages, voir Grosh et Glewwe (2000) et ONU (2005). Dal Poz et Gupta (2009) abordent certains problèmes spécifiques à la collecte des données dans le secteur de la santé.
3. À ce stade, la discussion peut s'appliquer à n'importe quelle population : l'ensemble de la population à l'étude, la population du groupe de traitement ou la population du groupe de comparaison.

4. Dans ce contexte, le terme « population » ne fait pas référence à la population d'un pays, mais plutôt à l'ensemble du groupe d'enfants qui nous intéresse, à savoir la « population à l'étude ».
5. Cette intuition est formalisée par le « théorème limite central ». Pour un résultat $\bar{y}$, ce théorème énonce que la moyenne de l'échantillon constitue une estimation valide de la moyenne de la population. Par ailleurs, pour un échantillon de taille *n* et une variance de σ dans la population, la variance de la moyenne de l'échantillon est inversement proportionnelle à la taille de l'échantillon :

$$\text{var}(\bar{y}) = \frac{\sigma^2}{n}.$$

   Plus la taille de l'échantillon *n* augmente, plus la variance des estimations d'échantillon s'approche de zéro. Autrement dit, la moyenne est estimée avec plus de précision avec de grands échantillons qu'avec de petits échantillons.
6. L'allocation du traitement par grappe est souvent incontournable à cause de considérations sociales ou politiques qui rendent impossible l'assignation aléatoire à l'intérieur des grappes. Dans le contexte d'une évaluation d'impact, la création de grappes est souvent nécessaire en raison du risque de débordements ou de diffusion des bénéfices du programme entre les individus au sein des grappes.
7. Outre la puissance, il convient également de fixer un niveau de confiance établissant une probabilité acceptable d'erreur de type I, généralement 0,05 (ou 0,01 pour un niveau plus conservateur).
8. Si les calculs de puissance sont effectués à partir de l'enquête de référence, l'auto-corrélation des résultats au fil du temps doit également être prise en compte.
9. Par exemple, Spybrook et al. (2008) ont développé Optimal Design, un logiciel convivial permettant de réaliser des calculs de puissance.
10. Il est généralement souhaitable d'avoir des groupes de traitement et de comparaison de la même taille. En effet, pour un nombre donné d'observations dans un échantillon, la puissance est optimisée en allouant la moitié des observations au groupe de traitement et l'autre moitié au groupe de comparaison. Toutefois, les groupes de traitement et de comparaison ne doivent pas systématiquement être de la même taille. Informez votre statisticien de toute contrainte s'opposant à l'utilisation de deux groupes de même taille ou de toute raison justifiant l'utilisation de groupes de tailles inégales.
11. Les questions de non-réponse et d'attrition sont abordées au chapitre 12 de manière plus détaillée.
12. Dans le contexte de l'évaluation d'un programme, l'ensemble de la population à l'étude peut être assigné au groupe de traitement ou au groupe de comparaison. Cette section décrit de façon générale la façon de prélever un échantillon de la population à l'étude totale.
13. Si l'on procède à un échantillonnage par grappes et que la liste des unités au sein des grappes n'est plus d'actualité, il faut envisager la possibilité d'effectuer une énumération exhaustive des unités au sein de chaque grappe. Par exemple, si l'échantillon est prélevé au sein d'une communauté, l'agence chargée de la collecte des données peut commencer par dresser la liste de tous les ménages du village avant de réaliser l'enquête.

14. Voir Cochran (1977) ; Lohr (1999) ; Kish (1995) ; Thompson (2002) ou, pour une présentation très abordable, Kalton (1983) pour une description de l'échantillonnage (y compris d'autres méthodes comme l'échantillonnage systématique ou en plusieurs étapes) plus approfondie que les concepts de base abordés ici. Grosh et Muñoz (1996) ; Fink (2008) ; Iarossi (2006) ; et ONU (2005) formulent des conseils pratiques sur l'échantillonnage.
15. Au sens strict, les échantillons sont prélevés à partir de cadres d'échantillonnage. Nous partons de l'hypothèse selon laquelle ce cadre coïncide parfaitement avec la population.

Références

Cochran, William G. 1977. *Sampling Techniques*. 3e édition. New York : John Wiley.

Dal Poz, Mario et Neeru Gupta. 2009. « Assessment of Human Resources for Health Using Cross-National Comparison of Facility Surveys in Six Countries. » *Human Resources for Health* 7 : 22.

Fink, Arlene G. 2008. *How to Conduct Surveys: A Step by Step Guide*. 4e édition. Beverly Hills, CA : Sage Publications.

Galiani, Sebastian, Paul Gertler et Ernesto Schargrodsky. 2005. « Water for Life: The Impact of the Privatization of Water Services on Child Mortality. » *Journal of Political Economy* 113 (1) : 83–120.

Grosh, Margaret et Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington DC : Banque mondiale.

Grosh, Margaret et Juan Muñoz. 1996. « A Manual for Planning and Implementing the Living Standards Measurement Study Survey. » Document de travail LSMS 126, Banque mondiale, Washington, DC.

Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*. Washington DC : Banque mondiale.

Kalton, Graham. 1983. *Introduction to Survey Sampling*. Beverly Hills, CA : Sage Publications.

Kish, Leslie. 1995. *Survey Sampling*. New York : John Wiley.

Lohr, Sharon. 1999. *Sampling: Design and Analysis*. Pacific Grove, CA : Brooks Cole.

Pradhan, Menno et Laura B. Rawlings. 2002. « The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund. » *Étude économique de la Banque mondiale* 16 (2) : 275–95.

Rosenbaum, Paul. 2009. *Design of Observational Studies*. New York : Springer Series in Statistics.

Spybrook, Jessaca, Stephen Raudenbush, Xiaofeng Liu, Richard Congdon et Andrés Martinez. 2008. *Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software*. New York : William T. Grant Foundation.

Thompson, Steven K. 2002. *Sampling*. 2e édition. New York : John Wiley.

ONU (Organisation des Nations Unies). 2005. *Household Sample Surveys in Developing and Transition Countries*. New York : Organisation des Nations Unies.

CHAPITRE 12

Collecter des données

Au chapitre 11, nous avons décrit le type de données nécessaires pour mener une évaluation et suggéré que la plupart des évaluations nécessitaient la collecte de nouvelles données. Nous avons également vu comment déterminer la taille de l'échantillon nécessaire et la façon de procéder à l'échantillonnage. Dans le présent chapitre, nous allons passer en revue les étapes de la collecte de données. Ces étapes doivent être bien comprises pour que l'évaluation d'impact soit fondée sur des données de qualité qui ne la compromettent pas. Dans un premier temps, vous devrez faire appel à une société ou un organisme gouvernemental spécialisé dans la collecte de données. Vous devrez en parallèle commanditer l'élaboration d'un questionnaire approprié. L'agence qui collecte les données devra recruter et former du personnel de terrain et procéder à un pilotage du questionnaire. Après avoir effectué les modifications nécessaires, la société ou l'organisme pourra entamer le travail sur le terrain. Enfin, les données collectées devront être saisies et validées avant d'être exploitées.

Choisir une entité compétente pour collecter les données

Vous devrez choisir assez tôt dans l'organisme qui sera responsable de la collecte des données, et ce en considérant nombre de facteurs importants. Ce travail peut potentiellement être réalisé par :

- l'institution responsable de la mise en œuvre du programme ;

- une autre institution gouvernementale qui possède de l'expérience dans la collecte de données (comme l'agence nationale de statistique) ; ou
- une société ou un groupe indépendant spécialisé dans la collecte de données.

L'entité collectant les données doit dans tous les cas travailler en étroite collaboration avec l'*organisme responsable de l'exécution du programme*. Étant donné que les données de référence doivent être collectées avant que le programme ne soit lancé, cette collaboration est nécessaire pour s'assurer qu'aucun aspect du programme n'est mis en œuvre avant que les données n'aient été collectées. Lorsque des données de référence sont nécessaires au fonctionnement du programme (par exemple, pour produire un indice de ciblage dans le contexte d'une évaluation fondée sur un modèle de discontinuité de la régression), l'organisme de collecte des données doit être en mesure de les traiter rapidement et de les transférer à l'institution responsable du programme. Une collaboration étroite est également nécessaire pour déterminer le moment le plus propice à la collecte des données de l'enquête de suivi. Par exemple, si vous avez choisi une assignation aléatoire par phase, l'enquête de suivi devra être réalisée avant que le programme ne soit déployé au sein du groupe de comparaison afin d'éviter toute contamination.

Au moment de choisir l'entité qui collectera les données, il est essentiel de garder à l'esprit qu'il faut employer des procédures de collecte identiques pour les groupes de comparaison et de traitement. Souvent, l'organisme chargé de l'exécution du programme n'a de contact qu'avec le groupe de traitement et n'est donc pas le mieux placé pour collecter des données pour le groupe de comparaison. Il serait risqué d'utiliser deux agences de collecte de données différentes pour les groupes de traitement et de comparaison, car cela peut entraîner des divergences dans les résultats mesurés pour les deux groupes du simple fait de l'utilisation de procédures différentes. Si l'organisme d'exécution ne peut pas collecter des données de manière efficiente pour le groupe de traitement et le groupe de comparaison, mieux vaut envisager le recrutement d'un autre partenaire.

Dans certains contextes, il est également recommandé de confier la collecte de données à une agence indépendante pour assurer l'objectivité du travail. Les craintes d'une éventuelle partialité de l'organisme responsable de l'exécution du programme peuvent être infondées, mais la contribution d'un organisme n'ayant aucun intérêt dans les résultats de l'évaluation peut renforcer la crédibilité de l'évaluation.

Étant donné que la collecte de données comprend une série d'opérations complexes, il est recommandé de faire appel à un organisme expérimenté. Rares sont les organismes responsables de l'exécution des programmes qui possèdent l'expérience suffisante pour collecter des données pour de grands échantillons avec une qualité suffisante pour une évaluation d'impact. Dans la plupart des cas, vous devrez envisager de mandater une institution locale, comme le bureau national de statistique, ou une société ou un groupe indépendant spécialisé.

Mandater une institution locale comme le *bureau national de statistique* peut être l'occasion, pour l'organisme, de se familiariser avec les études d'évaluation d'impact et d'étendre ses expertises. Toutefois, les bureaux nationaux de statistique ne

possèdent pas toujours les capacités suffisantes pour entreprendre des missions supplémentaires en dehors de leurs activités régulières. Ils peuvent également ne pas avoir l'expérience nécessaire pour réaliser des enquêtes dans le cadre d'évaluation d'impact, par exemple la mise en place de procédures de suivi des individus dans le temps. Dans ce cas, il peut être plus pratique de faire appel à une *société* ou à un groupe spécialisé dans la collecte de données.

Il n'est pas impératif que la même entité collecte les données de référence et de suivi. Par exemple, pour l'évaluation d'impact d'un programme de formation dont la population cible est composée des personnes s'étant inscrites, l'institution chargée de la formation peut être responsable de la collecte des données de référence au moment de l'inscription des participants. Il est toutefois peu probable que cette même institution soit la mieux placée pour collecter les données de suivi pour les groupes de traitement et de comparaison. Dans ce contexte, il peut être avantageux de contracter séparément la collection des diverses rondes de collectes de données, tout en s'assurant qu'aucune information utile au suivi des ménages ou des individus ne soit perdue entre les rondes et que les mêmes procédures soient utilisées pour les enquêtes de référence et de suivi.

Pour déterminer l'organisme le mieux à même de collecter les données pour l'évaluation d'impact, il faut tenir compte de toute une série de facteurs, y compris l'expérience en collecte de données, la capacité à collaborer avec l'organisme responsable de l'exécution du programme, l'impartialité, les opportunités de renforcer les capacités locales, la faculté d'adaptation au contexte de l'évaluation d'impact, ainsi que la qualité probable des données collectées. Rédiger un cahier des charges et solliciter des propositions techniques et financières est un moyen efficace de déterminer l'organisme le mieux placé pour collecter des données de qualité.

Étant donné que les délais de réalisation du travail et la qualité des données sont des facteurs déterminants pour la fiabilité de l'évaluation d'impact, le contrat avec l'agence de collecte des données doit être rédigé avec prudence. La portée des travaux et des résultats attendus doit être décrite de manière très précise. Il est également recommandé d'introduire au sein des contrats des mesures incitatives associées à de clairs indicateurs de qualité. Par exemple, comme nous le verrons ci-après, le taux de non-réponse est un indicateur clé de la qualité des données. Afin d'encourager les agences de collecte de données à limiter le taux de non-réponse, le contrat peut par exemple stipuler le paiement d'un coût unitaire pour 90 % de l'échantillon, d'un coût unitaire supérieur pour les unités comprises entre 90 % et 95 % et d'un coût unitaire encore plus élevé pour les unités comprises entre 95 % et 100 %. Un contrat séparé peut aussi être conclu pour la phase de suivi des non-répondants.

Élaboration du questionnaire

Au moment de commanditer une collecte de données, vous devez définir des objectifs précis et donner des directives spécifiques sur le contenu de l'instrument ou du

questionnaire à utiliser. Les instruments de collecte de données doivent permettre d'obtenir toutes les informations nécessaires pour répondre à la question de politique sur laquelle porte l'évaluation d'impact.

Choix des indicateurs

Comme nous l'avons vu, des *indicateurs* doivent être mesurés tout au long de la chaîne de résultats, y compris des indicateurs de résultat final, des indicateurs de résultats intermédiaires, et des indicateurs de la mise en œuvre de l'intervention, des facteurs exogènes et des caractéristiques de contrôle.

Il est important de choisir avec prudence les indicateurs à mesurer afin de limiter les coûts de la collecte des données, de simplifier la tâche de l'agence de collecte et d'améliorer la qualité des données collectées en réduisant le temps requis des répondants. Collecter des informations non pertinentes ou peu susceptibles d'être utilisées est très coûteux. La rédaction à l'avance d'un plan d'analyse des données vous permettra d'établir des priorités et de définir les informations nécessaires.

Les données sur les indicateurs de résultats et sur les caractéristiques de contrôle doivent être collectées de la même manière pour l'*enquête de référence* que pour l'*enquête de suivi*. Il est très souhaitable de collecter des données de référence. Même si vous utilisez les méthodes de l'assignation aléatoire ou de discontinuité de la régression, pour lesquelles de simples différences pour les indicateurs mesurées après l'intervention fournissent en principe l'impact d'un programme, les données de référence sont indispensables pour vérifier si la méthode d'évaluation d'impact est appropriée (voir la liste de l'encadré 8.1 au chapitre 8). Disposer de données de référence est également une assurance si la sélection aléatoire ne fonctionne pas parfaitement et que la méthode de la double différence doit être utilisée à la place. Les données de référence sont également utiles pendant la phase d'analyse puisque les variables de contrôle contenues dans les données de référence peuvent contribuer à augmenter la puissance statistique ou vous permettre d'analyser si les impacts varient pour différents sous-groupes. Enfin, les données de référence peuvent servir à améliorer la conception du programme. Ainsi, elles permettent parfois d'analyser l'efficacité du ciblage ou fournissent des informations supplémentaires sur les bénéficiaires à l'organisme responsable de l'exécution du programme.

Mesure des indicateurs

Une fois que les données essentielles à collecter sont définies, l'étape suivante consiste à déterminer la façon dont vous allez mesurer ces indicateurs. La *mesure* est un art en soi, et mieux vaut la confier à l'agence mandatée pour collecter les données, à des spécialistes ou aux évaluateurs. Des ouvrages entiers sont consacrés à la meilleure façon de mesurer des indicateurs particuliers, comme notamment la meilleure manière de formuler les questions qui figurent dans les enquêtes menées

auprès des ménages (voir Grosh et Glewwe [2000] et ONU [2005])[1] ou les procédures détaillées à suivre pour collecter des résultats d'examen ou des indicateurs de santé. Si ces considérations peuvent sembler laborieuses, elles n'en sont pas moins essentielles. Nous énonçons ici quelques principes généraux qui vous guideront dans la supervision de la collecte de données.

Les indicateurs de résultat doivent, dans la mesure du possible, être conformes aux meilleures pratiques locales et internationales. Il est toujours utile de se pencher sur la façon dont les indicateurs à l'étude ont été mesurés dans des enquêtes antérieures, à la fois sur le plan local et international. L'utilisation des mêmes indicateurs (et des mêmes modules ou questions pour l'enquête) permet de garantir la comparabilité entre les données existantes et les données collectées pour l'évaluation d'impact. Si vous décidez de choisir un indicateur qui n'est pas parfaitement comparable ou qui n'est pas bien mesuré, vous limitez l'utilité des résultats de l'évaluation.

Tous les indicateurs doivent être mesurés exactement de la même façon pour toutes les unités du groupe de traitement et du groupe de comparaison. L'utilisation de méthodes de collecte différentes (par exemple une enquête téléphonique dans un cas et des entretiens en face à face dans l'autre) risque de générer un biais. Ce risque est également présent si vous collectez des données à des moments différents pour les deux groupes (par exemple si vous collectez les données du groupe de traitement pendant la saison des pluies et celles du groupe de comparaison pendant la saison sèche). C'est pourquoi les procédures utilisées pour mesurer un indicateur de résultat doivent être formulées de manière très précise. Le processus de collecte des données doit être exactement le même pour toutes les unités. Dans le questionnaire, chaque module associé au programme doit être introduit sans affecter l'ordre ou le contexte des réponses dans d'autres sections du questionnaire.

Formatage des questionnaires

Des réponses différentes peuvent être obtenues en posant une même question de manière légèrement différente. Par conséquent, le contexte et la formulation des questions doivent être les mêmes pour toutes les unités afin d'éviter tout biais dans les réponses. Glewwe (ONU 2005) formule six recommandations spécifiques sur le contenu des questionnaires d'enquêtes auprès des ménages. Ces recommandations s'appliquent aussi à la plupart des autres instruments de collecte de données :

1. Chaque question doit être rédigée dans son intégralité dans le questionnaire afin que l'enquêteur puisse réaliser son entretien en lisant chaque question mot pour mot.

2. Le questionnaire doit inclure des définitions précises de tous les concepts clés mentionnés dans l'enquête afin que l'enquêteur puisse y faire référence pendant l'entretien si nécessaire.

3. Chaque question doit être aussi courte et simple que possible et être rédigée dans des termes simples du quotidien.
4. Les questionnaires doivent être conçus de façon à ce que les réponses à presque toutes les questions soient précodées.
5. Le système de codage doit être le même pour toutes les questions.
6. L'enquête doit clairement indiquer les questions à sauter en fonction des réponses aux questions précédentes.

Une fois le questionnaire rédigé par la personne mandatée, il doit être présenté à une équipe de spécialistes. Toutes les personnes participant à l'évaluation (décideurs, chercheurs, analystes et collecteurs de données) doivent être consultées pour savoir si le questionnaire permettra d'obtenir toutes les informations nécessaires.

Pilotage du questionnaire

Il est important que le questionnaire fasse l'objet d'un pilotage sur le terrain avant d'être finalisé. La réalisation d'un *pilote* permet de tester son contenu, son formatage et la formulation des questions. Il est essentiel de procéder à un *pilotage complet du questionnaire sur le terrain* dans des conditions réelles afin de vérifier la durée d'administration et de s'assurer que son contenu est suffisamment cohérent et complet pour mesurer toutes les informations pertinentes. Le pilotage sur le terrain fait partie intégrante du travail de conception du questionnaire.

Travail de terrain

Même si vous engagez un partenaire externe pour la collecte des données, il est essentiel que vous compreniez toutes les étapes de ce processus pour pouvoir garantir que les *mécanismes de contrôle de qualité* et les mécanismes *incitatifs* appropriés sont en place. L'organisme de collecte des données doit coordonner les travaux d'un grand nombre d'intervenants, parmi lesquels les enquêteurs, les superviseurs, les coordonnateurs de terrain ainsi que le personnel d'appui logistique en plus d'une équipe de programmateurs, de superviseurs et d'opérateurs de saisie. Un *plan de travail* précis doit être mis en place pour coordonner le travail de toutes ces équipes ; le plan de travail constitue donc un produit important.

Dès le début, le plan de travail doit prévoir une séance de *formation* de l'équipe de collecte avant que la collecte ne commence. À ce titre, un *manuel de référence* doit être rédigé et utilisé tout au long du travail sur le terrain. La formation est

essentielle pour s'assurer que les données sont collectées de la même manière par tous les intervenants. Le processus de formation est également une bonne occasion pour repérer les meilleurs enquêteurs et effectuer un dernier test des instruments et des procédures dans des conditions réelles. Une fois l'échantillon prélevé, les instruments conçus et testés, et les équipes formées, la collecte des données peut commencer. Il est utile de veiller à ce que le plan du travail de terrain prévoie que chaque équipe collecte des données pour le même nombre d'unités de traitement et de comparaison.

Comme nous l'avons vu au chapitre 11, la qualité de l'échantillonnage dépend essentiellement de la qualité des données recueillies. Toutefois, de nombreuses *erreurs non liées à l'échantillonnage* peuvent survenir pendant la collecte de données. Dans le contexte d'une évaluation d'impact, cela est d'autant plus problématique si ces erreurs diffèrent entre les groupes de traitement et de comparaison.

Une *non-réponse* apparait s'il est impossible de collecter des données exhaustives pour certaines unités de l'échantillon. Les échantillons effectifs se limitent aux unités pour lesquelles des données peuvent être collectées, les unités qui choisissent de ne pas participer à une enquête peuvent rendre l'échantillon moins représentatif et créer un biais dans les résultats de l'évaluation. L'*attrition* est une forme courante de non-réponse. Elle se produit lorsque des unités quittent l'échantillon entre deux rondes de collecte de données, par exemple par manque de suivi des migrants.

La non-réponse et l'attrition sont particulièrement problématiques dans le contexte des évaluations d'impact, car elles peuvent créer des différences entre le groupe de traitement et le groupe de comparaison. Par exemple, l'attrition peut varier dans les deux groupes : lors de la collecte de données de suivi, le taux de réponse parmi les unités traitées pourra être supérieur à celui des unités de comparaison. Par exemple, ceci peut être dû au fait que les unités de comparaison sont déçues de ne pas avoir été sélectionnées pour le programme ou sont plus susceptibles de migrer. Un problème de non-réponse peut également survenir si un questionnaire n'est pas complet pour certaines unités.

Concept clé :
La non-réponse caractérise le manque des données pour certaines unités de l'échantillon prévu. La non-réponse peut entraîner un biais dans les résultats de l'évaluation.

L'*erreur de mesure* est un autre type de problème pouvant générer un biais si elle est systématique. Une erreur de mesure survient lorsqu'il existe une différence entre la valeur d'une caractéristique fournie par le sondé et sa véritable valeur (inconnue) (Kasprzyk 2005). Cette différence peut être due à la façon dont le questionnaire est formulé ou à la méthode de collecte des données choisie. Elle peut également survenir par la faute des enquêteurs ou des sondés.

La qualité d'une évaluation d'impact dépend directement de la qualité des données collectées. Toutes les parties prenantes doivent connaître les *normes de qualité* qui régissent la collecte de données ; il faut notamment insister sur l'importance de ces normes durant la formation des enquêteurs et dans les manuels de référence. Il est également essentiel de définir des procédures détaillées pour réduire le taux de non-réponse ou (si cela est jugé acceptable) remplacer les unités introuvables de l'échantillon prévu. L'agence de collecte de données doit parfaitement comprendre

Concept clé :
Les meilleures pratiques en matière d'évaluation d'impact visent à limiter le taux de non-réponse et d'attrition à 5 %.

quels sont les taux de non-réponse et d'attrition acceptables. Les meilleures pratiques en matière d'évaluation d'impact visent à limiter le taux de non-réponse et d'attrition à 5 %. Cet objectif n'est pas toujours réalisable au sein de populations très mobiles, mais il fournit toutefois une référence utile. Il arrive que, pour limiter le taux de non-réponse, les sondés se voient offrir une compensation. Dans tous les cas, le contrat avec l'agence de collecte des données doit prévoir des mesures incitatives claires, par exemple une rémunération supérieure si le taux de non-réponse est inférieur à 5 % ou tout autre taux jugé acceptable.

En parallèle, des procédures d'*assurance de la qualité* bien définies doivent être établies à toutes les étapes du processus de collecte de données : conception des procédures d'échantillonnage, formulation du questionnaire, étapes de préparation, collecte, saisie, nettoyage et stockage des données.

Les contrôles de qualité doivent être considérés comme une priorité pendant les travaux sur le terrain afin de limiter les erreurs de non-réponse pour chaque unité. Des procédures précises doivent être mises en place pour revisiter les unités qui n'ont fourni aucune information ou pour lesquelles les informations sont incomplètes. Le processus de contrôle de la qualité doit comporter plusieurs filtres en prévoyant par exemple que les enquêteurs, les superviseurs et, si nécessaire, les coordonnateurs de terrain vérifient les cas de non-réponse. Les questionnaires correspondant aux cas de non-réponses doivent être clairement codés et consignés. Une fois les données saisies, le taux final de non-réponse peut être établi en révisant le statut de toutes les unités de l'échantillon prévu.

Des contrôles de qualité doivent également être effectués si les données d'un questionnaire sont incomplètes. Là encore, le processus de contrôle de la qualité doit comporter plusieurs filtres. L'enquêteur est chargé de vérifier les données immédiatement après leur collecte. Le superviseur et le coordonnateur de terrain doivent effectuer ultérieurement des vérifications aléatoires.

Les contrôles visant à détecter les erreurs de mesure sont plus compliqués, mais eux aussi essentiels pour déterminer si les informations ont été collectées correctement. Des contrôles de cohérence peuvent être intégrés au questionnaire. Par ailleurs, les superviseurs doivent effectuer des *vérifications ponctuelles* et des contre-vérifications pour s'assurer que les enquêteurs collectent les données conformément aux normes établies. Les coordonnateurs de terrain doivent également participer à ces contrôles pour réduire le risque de conflits d'intérêts au sein de la société de sondage.

Il est essentiel que toutes les étapes du contrôle de la qualité soient rendues explicites pour l'organisme chargé de la collecte des données. Vous pouvez également envisager de faire appel à un organisme indépendant pour superviser la qualité des activités de collecte de données. Ceci permet de limiter de façon significative les problèmes pouvant survenir en raison d'une supervision insuffisante de l'équipe de collecte de données.

Saisie et validation des données

Les enquêtes auprès des ménages sont généralement réalisées à l'aide d'un questionnaire papier bien que des instruments de collecte de données électroniques comme les ordinateurs portables et autres dispositifs portatifs deviennent plus courants. Dans tous les cas, les données doivent être numérisées et traitées. Un *logiciel de saisie de données* doit être créé et un système doit être mis en place pour gérer le flux des questionnaires à numériser. Il faut établir des normes et des procédures, et former les opérateurs de saisie, qui doivent tous suivre le même processus de saisie. Dans la mesure du possible, la saisie des données doit être intégrée aux opérations de collecte de données (y compris pendant la phase de pilotage) pour que tout problème concernant les données collectées puisse être rapidement identifié et immédiatement vérifié sur le terrain.

Si les enquêtes sont réalisées sur papier, il est impératif que les données brutes collectées soient saisies telles quelles, sans aucune modification. Afin de réduire les erreurs de saisie, il est recommandé d'exiger une procédure de *saisie de données en double aveugle* afin de repérer et de corriger toute erreur éventuelle.

Outre les contrôles de qualité effectués au cours du processus de saisie des données, le logiciel peut être programmé pour effectuer des vérifications automatiques d'erreurs non liées à l'échantillonnage (par exemple de non-réponse partielle et incohérences) susceptibles d'avoir été commises sur le terrain. Si le processus de saisie des données est intégré aux procédures de travail sur le terrain, les données incomplètes ou incohérentes peuvent être transmises aux enquêteurs pour leur vérification sur le terrain (Muñoz 2005, chapitre 15). Ce type d'intégration n'est pas sans poser de défis au niveau du flux organisationnel des opérations sur le terrain, mais il peut générer d'importants gains de qualité en réduisant les erreurs de mesure et en accroissant la puissance de l'évaluation d'impact. Le recours à une approche intégrée de ce type doit être envisagé au moment de la planification de la collecte de données. Les nouvelles technologies peuvent faciliter cette intégration.

Comme nous l'avons vu, la collecte de données implique une série d'opérations dont la complexité ne doit pas être sous-estimée. L'encadré 12.1 illustre le processus de collecte des données en vue de l'évaluation des programmes pilotes Atención a Crisis au Nicaragua, qui a généré des données de qualité avec un très faible taux d'attrition et de non-réponse tout en minimisant les erreurs de mesure et de saisie. Seule la mise en place de procédures et de mesures incitatives appropriées dès l'engagement d'un organisme de la collecte des données permet d'obtenir des données de qualité.

À l'issue du processus de collecte des données, les données doivent être transmises, accompagnées d'une documentation détaillée, comprenant un manuel et un dictionnaire complets, et stockées de façon sécurisée. Si les données sont collectées dans le cadre d'une évaluation d'impact, elles doivent également être accompagnées d'informations supplémentaires sur le traitement et la participation au programme de chaque unité. L'analyse d'évaluation d'impact sera d'autant plus rapide qu'elle pourra se reposer sur des données et une documentation complète, permettant ainsi son utilisation plus rapidement dans le cycle d'élaboration de politiques. Cela facilitera également le partage des informations.

Encadré 12.1 : Collecte de données pour l'évaluation des programmes pilotes Atención a Crisis au Nicaragua

En 2005, le Gouvernement du Nicaragua lance le programme pilote Atención a Crisis. L'objectif est d'évaluer l'impact de la combinaison d'un programme de transferts monétaires conditionnels (TMC) et de transferts productifs, tels que des transferts pour des investissements dans des activités non agricoles ou la participation à des formations professionnelles. Le projet pilote est mis en œuvre par le ministère de la Famille avec le soutien de la Banque mondiale.

Une assignation aléatoire en deux étapes est utilisée pour l'évaluation. Dans un premier temps, 106 communautés cibles sont réparties de manière aléatoire entre le groupe de comparaison et le groupe de traitement. Dans un second temps, au sein des communautés traitées, les ménages éligibles sont sélectionnés de manière aléatoire pour recevoir trois types de prestations : 1) un transfert monétaire conditionnel ; 2) un TMC plus une bourse permettant à l'un des membres du ménage de choisir une formation professionnelle ; et 3) un TMC plus un transfert pour permettre un investissement productif dans une activité non agricole, dans le but de créer des actifs et de diversifier les revenus (Macours et Vakis 2009).

Une enquête de référence est réalisée en 2005, avec une première enquête de suivi en 2006 et une deuxième enquête de suivi en 2008, deux ans après la fin de l'intervention. Des contrôles de qualité rigoureux sont mis en place à toutes les étapes du processus de collecte des données. Premièrement, les questionnaires sont testés sur le terrain et les enquêteurs sont formés à la fois dans des conditions théoriques et pratiques. Deuxièmement, un système de supervision sur le terrain est mis en place afin que tous les questionnaires soient révisés plusieurs fois par les enquêteurs, les superviseurs, les coordonnateurs de terrain et d'autres examinateurs. Troisièmement, un système de saisie des données en double aveugle est utilisé avec un programme complet de contrôle de la qualité capable de repérer les questionnaires incomplets ou incohérents. Les questionnaires présentant des non-réponses ou des incohérences sont systématiquement renvoyés sur le terrain pour vérification. Ces procédures et exigences sont décrites avec précision dans les termes de référence de l'agence de collecte des données.

Par ailleurs, des procédures de suivi détaillées sont mises en place pour limiter l'attrition. Au début, un recensement complet des ménages résidant dans les communautés de traitement et de comparaison en 2008 est entrepris en collaboration étroite avec les dirigeants communautaires. Au vu de l'importante mobilité géographique de la population, des mesures incitatives sont mises en place pour encourager la société de collecte de données à suivre les migrants dans tout le pays. Grâce à cette initiative, seulement 2 % des 4 359 ménages d'origine ne sont pas interrogés en 2009. La société de collecte de données est également mandatée pour suivre tous les individus des ménages interrogés en 2005. Là encore, seuls 2 % des individus auxquels les transferts du programme s'adressaient ne sont pas suivis (2 % étant par ailleurs décédés). Le taux d'attrition s'établit à 3 % pour tous les enfants des ménages interrogés en 2005 et à 5 % pour tous les individus des ménages interrogés en 2005.

Les taux d'attrition et de non-réponse donnent une bonne indication de la qualité de l'enquête. La société de collecte des données a déployé d'importants efforts et mis en place des mesures incitatives pour obtenir ces résultats remarquables. Il convient également de mentionner que le coût unitaire par ménage ou individu suivi est également beaucoup plus élevé. De plus, les contrôles de qualité rigoureux entraînent une augmentation des coûts et un allongement des délais de collecte des données. Toutefois, dans le contexte du projet pilote Atención a Crisis, l'échantillon reste représentatif à la fois au niveau des ménages et des individus plus de quatre ans après l'enquête de référence, l'erreur de mesure est minimisée et la fiabilité de l'évaluation est renforcée. Tous ces éléments font du programme Atención a Crisis l'un des projets de protection sociale dont la fiabilité peut être étudiée avec le plus de confiance.

Source : Macours et Vakis 2009 ; auteurs.

Note

1. Voir également Fink et Kosecoff (2008) ; Iarossi (2006) ; et Leeuw, Hox et Dillman (2008), qui fournissent de nombreux conseils pratiques sur la collecte de données.

Références

Fink, Arlene G. et Jacqueline Kosecoff. 2008. *How to Conduct Surveys: A Step by Step Guide*. 4e édition. Londres : Sage Publications.

Glewwe, Paul. 2005. « An Overview of Questionnaire Design for Household Surveys in Developing Countries. » In *Household Sample Surveys in Developing and Transition Countries*, chapitre 3. New York : Organisation des Nations Unies.

Grosh, Margaret et Paul Glewwe, eds. 2000. *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington DC : Banque mondiale.

Iarossi, Giuseppe. 2006. *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents*. Washington DC : Banque mondiale.

Kasprzyk, Daniel. 2005. « Measurement Error in Household Surveys: Sources and Measurement. » In *Household Sample Surveys in Developing and Transition Countries*, chapitre 9. New York : Organisation des Nations Unies.

Leeuw, Edith, Joop Hox et Don Dillman. 2008. *International Handbook of Survey Methodology*. New York : Taylor & Francis Group.

Macours, Karen et Renos Vakis. 2009. « Changing Household Investments and Aspirations through Social Interactions: Evidence from a Randomized Experiment. » Document de travail consacré à la recherche sur les politiques 5137, Banque mondiale, Washington, DC.

Muñoz, Juan. 2005. « A Guide for Data Management of Household Surveys. » In *Household Sample Surveys in Developing and Transition Countries*, chapitre 15. New York : Organisation des Nations Unies.

ONU (Organisation des Nations Unies). 2005. *Household Sample Surveys in Developing and Transition Countries*. New York : Organisation des Nations Unies.

CHAPITRE 13

Production et diffusion des résultats

Dans ce chapitre, nous abordons le contenu et la diffusion des divers rapports produits au cours d'une évaluation d'impact. Pendant la phase de préparation, le gestionnaire de l'évaluation commence par élaborer un plan de réalisation de l'évaluation d'impact qui détaille les objectifs, la méthode, les stratégies d'échantillonnage et de collecte de données pour l'évaluation (l'encadré 13.1 propose une ébauche du plan d'évaluation). Les différents éléments du plan d'évaluation sont présentés dans les chapitres 1 à 12 ci-dessus.

Une fois l'évaluation en cours, les évaluateurs produisent plusieurs rapports, dont un *rapport de référence*, au moins un *rapport d'évaluation d'impact* et *des notes de synthèse politique*. Les évaluateurs fournissent également des bases de données documentées. Lorsque le rapport d'évaluation d'impact est terminé et que les résultats sont connus, il faut déterminer la meilleure façon de diffuser les conclusions auprès des décideurs et autres parties prenantes concernées. Le présent chapitre est consacré à la production et à la diffusion des résultats de l'évaluation d'impact.

Les produits de l'évaluation

Les principaux produits d'une évaluation sont le rapport d'évaluation d'impact et des notes de synthèse politique résumant les principales conclusions. La réalisa-

Encadré 13.1 : Exemple de structure d'un plan d'évaluation d'impact

1. Introduction
2. Description de l'intervention
3. Objectifs de l'évaluation
   - 3.1 Hypothèses, théorie du changement, chaîne de résultats
   - 3.2 Questions de politique
   - 3.3 Indicateurs de résultat clés
4. Méthode d'évaluation
5. Échantillonnage et données
   - 5.1 Stratégie d'échantillonnage
   - 5.2 Calculs de puissance
6. Plan de collecte des données
   - 6.1 Enquête de référence
   - 6.2 Enquête(s) de suivi
7. Produits
   - 7.1 Rapport de référence
   - 7.2 Rapport d'évaluation d'impact
   - 7.3 Note de synthèse politique
   - 7.4 Bases de données documentées
8. Plan de diffusion
9. Questions éthiques
10. Calendrier
11. Budget et financement
12. Composition de l'équipe d'évaluation

tion d'un rapport final d'évaluation peut prendre plusieurs années puisque les conclusions ne peuvent être obtenues que lorsque toutes les données de suivi ont été collectées. En raison de ce délai, les décideurs demandent souvent à recevoir des rapports d'évaluation intermédiaires, comme un rapport de référence, afin de disposer d'informations préliminaires pour alimenter le dialogue et les décisions de politique publique[1].

Comme nous l'avons vu au chapitre 10, le gestionnaire de l'évaluation travaille en collaboration avec des analystes pour produire le rapport de référence et le rapport final. Les analystes sont des experts en statistique ou en économétrie qui peuvent

réaliser l'analyse de l'évaluation d'impact au moyen d'un logiciel statistique comme Stata, SPSS ou R. Ils sont chargés de garantir la qualité, la rigueur scientifique et la crédibilité des résultats. Nous n'abordons pas dans ce chapitre la façon d'analyser les données[2], mais plutôt le contenu des rapports produits à partir des données.

Produit intermédiaire : le rapport de référence

Le principal objectif du *rapport de référence* est de déterminer si la méthode d'évaluation d'impact choisie est valide dans la pratique et de décrire les caractéristiques de la population éligible et les indicateurs de résultats de référence (avant le programme). Le rapport de référence contient également des informations sur le programme et ses bénéficiaires qui peuvent être utiles pour améliorer à la fois la mise en œuvre du programme et son évaluation. L'encadré 13.2 présente un exemple du contenu d'un rapport de référence[3].

Le rapport de référence est produit à partir de l'analyse de données de base et de données administratives décrivant les unités qui font partie du groupe de traitement ou de comparaison. L'assignation de ménages, d'individus ou d'établissements au

Encadré 13.2 : Exemple de structure d'un rapport de référence

1. Introduction
2. Description de l'intervention (bénéfices, règles d'éligibilité, etc.)
3. Objectifs de l'évaluation
   - 3.1 Hypothèses, théorie du changement, chaîne de résultats
   - 3.2 Questions de politique
   - 3.3 Indicateurs de résultat clés
4. Méthode d'évaluation
   - 4.1 Méthode prévue initialement
   - 4.2 Participants et non participants effectifs au programme
5. Échantillonnage et données
   - 5.1 Stratégie d'échantillonnage
   - 5.2 Calculs de puissance
   - 5.3 Données collectées
6. Validation de la méthode d'évaluation
7. Statistiques descriptives complètes
8. Conclusion et recommandations pour la mise en œuvre du programme

groupe de traitement ou au groupe de comparaison s'effectue généralement après la collecte des données de référence. Par conséquent, l'assignation de chaque unité au groupe de traitement ou de comparaison est souvent enregistrée dans une base de données administrative distincte. Par exemple, un tirage au sort peut être organisé pour déterminer les communautés qui bénéficieront d'un programme de transferts monétaires parmi toutes les communautés éligibles auprès desquelles l'enquête de référence a été réalisée. Dans ce cas, les analystes doivent effectuer un croisement des données administratives et des données de référence. Si l'évaluation porte sur, disons, plus de 100 unités éligibles, il ne sera pas pratique d'effectuer un croisement par nom des données de référence et des données administratives. Il faudra attribuer à chaque unité éligible un numéro ou un *identifiant* unique qui servira à l'identifier dans toutes les sources de données, y compris dans les bases de données de référence et administratives.

Les premières sections du rapport de référence approfondissent le plan d'évaluation d'impact en présentant le contexte de l'évaluation, le contenu de l'intervention (bénéfices du programme et règles d'assignation), les objectifs de l'évaluation (théorie du changement, principales questions de politique, hypothèses et indicateurs) et la méthode choisie pour l'évaluation. La section consacrée à la conception de l'évaluation doit déterminer si l'assignation des bénéfices du programme a été conforme à la méthode prévue. Étant donné que l'assignation est généralement réalisée juste après l'enquête de référence, il est recommandé de présenter des informations sur l'assignation effective dans le rapport de référence. La section sur l'échantillonnage commence généralement par une description de la stratégie d'échantillonnage et des calculs de puissance effectués avant de passer aux détails sur la façon dont les données de référence ont été collectées et le type d'informations qui sont disponibles. Le rapport doit mentionner toutes les éventuelles difficultés rencontrées lors de la collecte des données de référence et présenter des indicateurs de la qualité des données, par exemple les taux de non-réponse. À ce titre, le rapport de référence peut mettre en évidence les principaux problèmes à résoudre au moment de collecter les données du suivi. Par exemple, si le taux de non-réponse est élevé lors de l'enquête de référence, les évaluateurs devront penser à élaborer de nouvelles procédures pour s'assurer que cela ne se reproduise pas pour l'enquête de suivi.

Comme nous l'avons mentionné, le principal objectif du rapport de référence est de juger si la méthode d'évaluation choisie et présentée dans le plan d'évaluation reste valable en pratique. Nous avons vu au chapitre 8 que la plupart des méthodes d'évaluation d'impact ne produisent des estimations valides du contrefactuel que dans le cadre d'hypothèses spécifiques. L'encadré 8.1 (chapitre 8) présente la liste des tests qui peuvent servir à tester la pertinence d'une méthode en fonction du contexte. Certains de ces tests ne nécessitent pas de données de suivi et peuvent être appliqués dès que les données de référence sont disponibles. Par exemple, si la méthode de l'assignation aléatoire ou de l'offre aléatoire est utilisée, le rapport de référence doit préciser si les groupes de traitement et de comparaison présentent les mêmes caractéristiques. Si l'évaluation est fondée sur la méthode de la discontinuité

de la régression, le rapport de référence doit considérer si l'indice d'éligibilité est continu autour du seuil d'éligibilité. Même si ces tests de falsification ne garantissent pas que le groupe de comparaison reste valide jusqu'à l'enquête de suivi, il est impératif de les présenter dans le rapport de référence.

En plus d'éprouver la validité de la méthode d'évaluation, le rapport de référence doit comporter des tableaux décrivant les caractéristiques de l'échantillon d'évaluation. Ces tableaux peuvent faciliter la mise en œuvre du programme en permettant aux gestionnaires de mieux cerner le profil des bénéficiaires et d'adapter l'intervention à leurs besoins. Par exemple, les gestionnaires peuvent adapter le contenu des formations proposées par un programme de formation des jeunes en ayant une meilleure idée du niveau d'éducation ou de l'expérience professionnelle moyenne des participants.

Du point de vue de l'évaluation, l'enquête de référence génère souvent des informations qui n'étaient pas disponibles au moment de la formulation du plan d'évaluation. Supposons que vous cherchiez à évaluer l'impact d'un programme de santé dans les villages sur l'incidence de la diarrhée chez les enfants. Au moment de la rédaction du plan d'évaluation, il se peut que vous ne connaissiez pas le taux d'incidence exact de la diarrhée. Votre plan d'évaluation contient seulement une estimation sur laquelle sont fondés les calculs de puissance. Cependant, une fois que vous disposez des données de référence, vous pouvez vérifier le taux d'incidence de la diarrhée et vérifier si la taille initiale de votre échantillon est adéquate. Si vous constatez que les valeurs de référence des indicateurs de résultat sont différentes de celles utilisées pour les calculs de puissance initiaux, le rapport de référence pourra actualiser les calculs de puissance.

Afin de garantir la crédibilité des résultats finaux de l'évaluation, il est judicieux de demander à des spécialistes externes d'effectuer une revue critique du rapport de référence. La diffusion du rapport de référence peut également renforcer le dialogue politique entre les parties prenantes au cours du cycle d'évaluation.

Produits finaux : rapport d'évaluation d'impact, note de synthèse politique et bases de données

Le *rapport final d'évaluation d'impact* est le principal produit de l'évaluation. Il est rédigé à partir des données de suivi[4]. Le principal objectif du rapport d'évaluation est de présenter les résultats de l'évaluation et de répondre à toutes les questions de politique posées initialement. Par ailleurs, le rapport doit montrer que l'évaluation est fondée sur des estimations valides du contrefactuel et que les impacts identifiés sont entièrement attribuables au programme.

Le rapport d'évaluation d'impact final est un rapport exhaustif qui résume l'ensemble des travaux accomplis dans le cadre de l'évaluation et qui inclut une description détaillée de l'analyse des données et des spécifications économétriques ainsi qu'une analyse des résultats, des tableaux et des annexes. L'encadré 13.3 présente un exemple de contenu d'un rapport d'évaluation d'impact. Il existe de nombreux bons

exemples de rapports d'évaluation d'impact, comme Maluccio et Flores (2005), Levy et Ohls (2007) ou Skoufias (2005) pour les programmes de transferts monétaires conditionnels ; Card et al. (2007) pour un programme de formation des jeunes ; Cattaneo et al. (2009) pour un programme de logement ; et Basinga et al. (2010) pour un programme de paiement à la performance dans le secteur de la santé.

Comme pour le rapport de référence, les évaluateurs et analystes collaborent pour produire le rapport final d'évaluation d'impact. Ils commencent par produire une base de données contenant les données de référence, les données de suivi et les données administratives sur la mise en œuvre du programme, ainsi que les données sur l'assignation initiale aux groupes de traitement et de comparaison. Toutes ces sources de données doivent être croisées et consolidées en utilisant l'identifiant unique de chaque unité.

Étant donné que le rapport final d'évaluation d'impact est le principal produit de l'évaluation, il doit passer en revue les informations clés du plan d'évaluation et du rapport de référence avant de passer à l'analyse des résultats. La section d'introduction du rapport final doit présenter la motivation pour l'intervention et l'évaluation

Encadré 13.3 : Exemple de structure d'un rapport d'évaluation

1. Introduction
2. Description de l'intervention (bénéfices, règles d'éligibilité, etc.)
   - 2.1. Conception
   - 2.2 Mise en œuvre
3. Objectifs de l'évaluation
   - 3.1 Hypothèses, théorie du changement, chaîne de résultats
   - 3.2 Questions de politique
   - 3.3 Indicateurs de résultat clés
4. Méthode d'évaluation
   - 4.1 Théorie
   - 4.2 Pratique
5. Échantillonnage et données
   - 5.1 Stratégie d'échantillonnage
   - 5.2 Calculs de puissance
   - 5.3 Données collectées
6. Validation de la méthode d'évaluation
7. Résultats
8. Tests de sensibilité
9. Conclusion et recommandations de politique

puis décrire l'intervention (bénéfices et règles d'assignation), les objectifs de l'évaluation (théorie du changement, principales questions de politique, hypothèses et indicateurs), la méthode d'évaluation et la façon dont elle a été mise en œuvre.

En général, l'interprétation des résultats dépend de la façon dont l'intervention a été mise en œuvre. Le rapport d'évaluation final doit donc aborder en détail la façon dont l'intervention a été mise en œuvre. Ces informations peuvent être présentées avant les résultats, par exemple en décrivant les données sur la mise en œuvre du programme obtenues à partir des enquêtes de suivi ou de sources administratives complémentaires.

La section sur l'échantillonnage et les données doit contenir une description de la stratégie d'échantillonnage et des calculs de puissance avant l'analyse détaillée des données de référence et de suivi. Les indicateurs clés de qualité des données, comme les taux de non-réponse et d'attrition, doivent être présentés pour chaque ronde de données. Si ces taux sont élevés, l'analyste doit expliquer dans quelle mesure ils peuvent affecter l'interprétation des résultats. Par exemple, il est essentiel de vérifier si les niveaux d'attrition ou de non-réponse sont similaires dans les groupes de comparaison et de traitement.

Une fois les données décrites, le rapport peut présenter les résultats pour chaque question de politique ainsi que pour tous les indicateurs de résultat identifiés dans les objectifs de l'évaluation. La structure de la présentation des résultats dépend du type de questions de politique à l'étude. Par exemple, l'évaluation vise-t-elle à éprouver la validité de différentes alternatives de conception de programme ou seulement l'efficacité d'une intervention ? Cela intéresse-t-il les décideurs de savoir si les impacts du programme varient entre différents sous-groupes ? Pour les évaluations bien conçues et bien mises en œuvre, des résultats rigoureux peuvent être présentés de manière intuitive.

Comme nous l'avons mentionné, le rapport d'évaluation d'impact doit établir que les impacts estimés sont entièrement attribuables au programme. Il doit donc comporter une étude approfondie de la validité de la méthode d'évaluation, en commençant par présenter les résultats des tests de falsification effectués avec les données de référence (encadré 8.1, chapitre 8), puis des tests éventuellement effectués à partir des données de suivi. Par exemple, si la méthode de la double différence est choisie, certains des tests de falsification décrits dans l'encadré 8.1 ne peuvent être effectués que si les données de suivi sont disponibles.

L'introduction du rapport d'évaluation doit énumérer toute difficulté rencontrée par la méthode d'évaluation entre l'enquête de référence et l'enquête de suivi. Par exemple, le manque d'adhérence des participants au groupe de traitement ou au groupe de comparaison a des implications importantes au niveau de l'analyse et de l'interprétation des résultats et doit donc être mentionné dès le début du rapport.

Le rapport doit également contenir des renseignements sur le nombre d'unités assignées au groupe de traitement n'ayant pas bénéficié du programme et sur le nombre d'unités assignées au groupe de comparaison en ayant bénéficié. L'analyse doit être ajustée pour prendre en compte toute différence observée par rapport à l'assignation initiale (ces techniques sont décrites dans la partie 2).

En parallèle aux tests sur la validité de la méthode d'évaluation, le rapport final doit fournir une analyse approfondie de la nature, de la fiabilité et de la sensibilité des résultats. Il doit contenir une série de tests de sensibilité portant sur la méthodologie d'évaluation employée. Par exemple, si une méthode d'appariement est utilisée, le rapport doit présenter les résultats de plusieurs techniques d'appariement alternatives. Les analystes ont la responsabilité de déterminer et de présenter les tests de robustesse nécessaires à l'évaluation. La dernière partie du rapport doit fournir une réponse claire à toutes les questions de politique motivant l'évaluation et présenter des recommandations de politique détaillées fondées sur les résultats.

Il est particulièrement important de comprendre comment l'intervention a été mise en œuvre si les résultats de l'évaluation font état d'un impact limité ou négatif. L'absence de résultats ou des résultats négatifs ne justifient pas des sanctions à l'encontre du programme ou des évaluateurs. Au contraire, ils constituent une occasion d'expliquer clairement ce qui n'a pas fonctionné comme prévu, un élément essentiel pour améliorer les programmes et politiques. Lorsque des signes indiquent que l'évaluation va produire des résultats nuls ou négatifs, il est particulièrement important que l'équipe d'évaluation communique continuellement avec les décideurs et responsables du programme. Des évaluations de processus ou des travaux qualitatifs complémentaires peuvent contribuer à expliquer la raison pour laquelle un programme n'a pas produit les résultats escomptés. Une absence de résultats causée par la mise en œuvre imparfaite du programme doit être différenciée d'une absence de résultats causée par un programme bien mis en œuvre, mais mal conçu[5]. En général, les évaluations d'alternatives de conception d'un même programme sont les plus utiles pour distinguer formellement les caractéristiques qui fonctionnent ou pas.

Globalement, l'analyse finale des données doit générer des preuves convaincantes que les impacts détectés sont effectivement attribuables au programme. Pour garantir l'objectivité et la légitimité des résultats, tous les rapports doivent faire l'objet d'une revue critique externe et de consultations techniques rigoureuses avant d'être finalisés. Le contenu du rapport final d'évaluation d'impact peut par la suite être retravaillé et publié dans un journal académique plus technique, renforçant ainsi la crédibilité des résultats de l'évaluation.

Outre le rapport d'évaluation complet, les évaluateurs doivent produire une ou plusieurs notes de synthèse politique pour communiquer les résultats aux décideurs et aux autres parties prenantes. La note de synthèse politique présente les principales conclusions de l'évaluation sous forme de graphiques, de diagrammes ou d'autres formats lisibles, et résume les recommandations de politique de l'analyse. Elle contient également un résumé des caractéristiques techniques de l'évaluation. Elle peut être rendue publique en format papier ou mise en ligne et diffusée aux politiciens, à la société civile et aux médias. De bons exemples de notes de synthèse politique se trouvent sur les sites Internet de Poverty Action Lab (JPAL) ou du Réseau de développement humain de la Banque mondiale (par exemple, Poverty Action Lab 2008 ; Réseau de développement humain de la Banque mondiale 2010).

Une base de données documentée constitue le dernier produit majeur généré par une évaluation d'impact. La documentation peut être effectuée à l'aide d'outils comme le Microdata Management Toolkit de l'International Household Survey Network (http://www.ihsn.org). Les décideurs et les évaluateurs conviennent généralement d'un calendrier pour la réalisation de l'analyse et le partage des données d'évaluation. Il est important de mettre les données à la disposition du public pour assurer la transparence de l'évaluation. Ainsi, les résultats peuvent aussi être vérifiés et validés par des chercheurs indépendants. La diffusion publique des données encourage d'autres chercheurs à effectuer des analyses supplémentaires, ce qui peut générer de nouvelles informations et de nouveaux résultats pertinents pour le programme. Lorsque les données sont rendues publiques, il est important de garantir l'anonymat de tous les sujets étudiés. Toute information permettant d'identifier les sondés (nom, adresse ou informations sur le lieu) doit être supprimée des bases de données publiées. Les renseignements personnels doivent être traités de manière confidentielle et ne doivent servir que dans le cadre de nouvelles activités de collecte de données dument autorisées.

Diffusion des résultats

Au-delà de la simple production des résultats, l'objectif des évaluations d'impact est de renforcer l'efficacité des politiques publiques et de contribuer à améliorer le bien-être des populations. Afin de garantir que l'évaluation d'impact est prise en compte dans les décisions de politique, il est essentiel d'établir une communication claire entre toutes les parties prenantes (décideurs, société civile et médias). Les évaluations influentes comprennent souvent un plan de diffusion détaillé qui décrit la façon dont les parties prenantes doivent être informées et mobilisées tout au long du cycle d'évaluation. Ce plan de diffusion peut faciliter la prise en compte des conclusions par les décideurs et garantir que l'évaluation d'impact produise de véritables résultats.

Dès les premières phases de la conception de l'évaluation, les évaluateurs peuvent établir de solides canaux de communication avec les décideurs. Comme nous

l'avons souligné dans notre présentation des méthodes d'évaluation, la conception de l'évaluation dépend directement de la conception et du mode de mise en œuvre du programme. Il est donc essentiel que les évaluateurs externes et les décideurs qui commandent l'évaluation collaborent étroitement pendant la phase de conception du programme. Si l'équipe d'évaluation est bien organisée, il sera plus facile de faire en sorte que l'évaluation réponde aux besoins des décideurs et les progrès et les résultats seront régulièrement communiqués à ces derniers.

Le *plan de diffusion* doit énoncer comment l'équipe d'évaluation contribuera à soutenir la demande pour les résultats de l'évaluation et assurer leur utilisation dans les prises de décisions. Les évaluateurs doivent sensibiliser toutes les parties prenantes internes et externes en leur communiquant efficacement les résultats tout au long du cycle d'évaluation. Au moment de lancer l'évaluation, l'organisation d'un atelier préalable avec les responsables du programme et les principales parties prenantes peut permettre d'établir un consensus sur les objectifs principaux, les questions de politique clés et la méthode de l'évaluation. C'est également l'occasion de mener des consultations et d'assurer que l'évaluation répond parfaitement aux besoins des parties prenantes, en plus de les sensibiliser à l'évaluation et de renforcer leur intérêt pour les résultats.

Pendant l'évaluation, des réunions régulières d'un comité interinstitutionnel ou une table ronde permanente peuvent garantir que les travaux de l'équipe d'évaluation restent pertinents. Ces fora peuvent permettre d'obtenir des commentaires et des réactions sur des produits tels que les termes de références, les instruments d'enquête, les modes de diffusion des résultats ou la meilleure manière d'atteindre les hauts responsables.

Il est important d'organiser des événements de diffusion pour les produits intermédiaires, comme le rapport de référence, afin d'entretenir un dialogue actif avec les utilisateurs de l'évaluation. Prévoir des consultations sur le rapport de référence permet à la fois de diffuser les résultats intermédiaires pertinents et de continuer à sensibiliser les parties prenantes sur les résultats à venir.

Avant de finaliser le rapport d'évaluation, certains évaluateurs décident d'organiser une dernière consultation pour donner aux parties prenantes la possibilité de commenter les résultats. Ces consultations peuvent contribuer à améliorer la qualité des résultats et leur acceptation. Une fois que le rapport final d'évaluation d'impact et les notes de synthèse politique sont prêts, des événements de diffusion peuvent être organisés pour communiquer les résultats à toutes les parties prenantes. Un atelier national de consultation et de diffusion réunissant un grand nombre de parties prenantes est un bon moyen de discuter des résultats, de recevoir des commentaires et de définir les changements de politique qui pourraient être entrepris sur la base des résultats. Cet atelier peut être suivi d'un atelier de diffusion destiné aux hauts responsables (voir encadré 13.4). En dehors du pays concerné, les résultats peuvent être diffusés à l'occasion de conférences, de séminaires ou d'autres rencontres s'ils sont jugés utiles pour les politiques d'autres pays. D'autres circuits de diffusion innovants, comme les interfaces en ligne, peuvent permettre de renforcer la visibilité des conclusions.

Encadré 13.4 : Diffuser les résultats d'une évaluation pour améliorer les politiques

L'évaluation d'une initiative de paiement à la performance des prestataires de santé au Rwanda donne un bon exemple de stratégie de diffusion efficace. Sous la direction du ministère de la Santé, une équipe composée d'universitaires locaux et de spécialistes de la Banque mondiale est chargée de mener l'évaluation. Diverses parties prenantes participent à l'évaluation dès son lancement, ce qui se révèle essentiel pour garantir son succès et un fort support politique tout au long de sa mise en œuvre. Les résultats finaux de l'évaluation (Basinga et al. 2010) sont présentés à l'occasion d'un atelier public d'une journée réunissant de hauts responsables et plusieurs parties prenantes. Grâce à ces canaux de communication, les conclusions influencent fortement la formulation de la politique de santé au Rwanda. Les résultats sont également diffusés dans des conférences internationales sur la santé et par le biais d'un site Internet.

Source : Morgan 2010.

Au final, la diffusion des résultats d'une évaluation d'impact conformément à un plan bien conçu couvrant tout le cycle d'évaluation est essentielle pour que les résultats influencent le dialogue politique. Les évaluations d'impact ne peuvent remplir leur objectif premier, à savoir l'amélioration de l'efficacité des programmes de développement, que si les résultats sont partagés avec les décideurs et utilisés dans le processus de prise de décision.

Notes

1. Une évaluation peut générer d'autres produits intermédiaires. Par exemple, des évaluations qualitatives ou de processus fournissent de précieuses informations complémentaires avant la rédaction du rapport d'évaluation d'impact final. Nous nous concentrons ici sur le rapport de référence, car il constitue le principal produit intermédiaire des évaluations d'impact quantitatives qui font l'objet de cet ouvrage.
2. Khandker et al. (2009) présentent une introduction à l'évaluation qui comprend une revue de l'analyse des données, y compris les commandes Stata correspondantes à chaque méthode d'évaluation d'impact.
3. Cette structure est indicative et peut être adaptée en fonction de la nature de chaque évaluation, par exemple en modifiant l'ordre ou le contenu des différentes sections.

4. Lorsque différentes rondes de données de suivi sont collectées, un rapport d'évaluation d'impact peut être rédigé pour chaque ronde, et les résultats peuvent être comparés pour déterminer comment les impacts du programme varient au fil du temps.
5. Comme nous l'avons vu au chapitre 1, c'est la raison pour laquelle les essais d'efficacité pilotes visant à limiter les problèmes de mise en œuvre sont utiles pour déterminer si un programme donné est efficace quand il se déroule dans des circonstances idéales. Une fois la validation de principe documentée, l'étude pilote peut être étendue pour être testée dans des conditions réelles.

Références

Basinga, Paulin, Paul J. Gertler, Agnes Binagwaho, Agnes L. B. Soucat, Jennifer R. Sturdy et Christel M. J. Vermeersch. 2010. « Paying Primary Health Care Centers for Performance in Rwanda. » Document de travail consacré à la recherche sur les politiques 5190, Banque mondiale, Washington, DC.

Card, David, Pablo Ibarraran, Ferdinando Regalia, David Rosas et Yuri Soares. 2007. « The Labor Market Impacts of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation: Evidence from a Randomized Evaluation. » NBER Working Paper 12883, National Bureau of Economic Research, Washington, DC.

Cattaneo, Matias, Sebastian Galiani, Paul Gertler, Sebastian Martinez et Rocio Titiunik. 2009. « Housing, Health and Happiness. » *American Economic Journal : Economic Policy* 1 (1) : 75–105.

Khandker, Shahidur R., Gayatri B. Koolwal et Hussain A. Samad. 2009. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington DC : Banque mondiale.

Levy, Dan et Jim Ohls. 2007. « Evaluation of Jamaica's PATH Program: Final Report. » Ref. No. 8966-090, Mathematica Policy Research, Inc., Washington, DC.

Maluccio, John et Rafael Flores. 2005. « Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Proteccion Social. » Rapport de recherche 141, Institut international de recherche sur les politiques alimentaires, Washington, DC.

Morgan, Lindsay. 2010. « Signed, Sealed, Delivered? Evidence from Rwanda on the Impact of Results-Based Financing for Health. » Note de synthèse politique HRBF, Banque mondiale, Washington, DC.

Poverty Action Lab. 2008. « Solving Absenteeism, Raising Test Scores. » Policy Briefcase 6. http://www.povertyactionlab.org.

Skoufias, Emmanuel. 2005. « PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico. » Rapport de recherche 139, Institut international de recherche sur les politiques alimentaires, Washington, DC.

Réseau de développement humain de la Banque mondiale. 2010. « Does Linking Teacher Pay to Student Performance Improve Results? » Notes de synthèse politique, série 1, Banque mondiale, Washington DC. http://www.worldbank.org/hdchiefeconomist.

CHAPITRE 14

Conclusion

Le présent ouvrage est un guide pratique sur la conception et la mise en œuvre des évaluations d'impact. Son contenu s'adresse à trois groupes de lecteurs : 1) les décideurs qui exploitent les informations générées par les évaluations d'impact, 2) les gestionnaires de projet et les professionnels du développement qui commanditent des évaluations, et 3) les techniciens qui conçoivent et mettent en œuvre des évaluations d'impact. L'évaluation d'impact vise essentiellement à générer des preuves quant à l'efficacité ou l'inefficacité des politiques sociales. Une évaluation d'impact classique compare les résultats en la présence et en l'absence d'un programme à l'étude. Les évaluations d'impact peuvent également permettre d'étudier différentes options de mise en œuvre d'un même programme ou de comparer les performances de différents programmes.

Les évaluations d'impact constituent, selon nous, un investissement justifié pour de nombreux programmes. Complétées par des méthodes de suivi et d'autres formes d'évaluation, elles permettent de mieux comprendre l'efficacité des politiques sociales. Nous avons présenté différentes méthodes d'évaluation d'impact ainsi que leurs avantages et leurs inconvénients en termes de mise en œuvre, d'économie politique, de contraintes financières et d'interprétation des résultats. Nous avons montré qu'une bonne méthode est une méthode qui s'adapte au contexte opérationnel et non le contraire. Enfin, nous avons formulé des conseils pratiques et passé en revue des outils qui visent à faciliter la conduite d'une évaluation et l'exploitation de ses résultats.

Les évaluations d'impact sont des entreprises complexes nécessitant la coordination de nombreux partenaires et activités. La liste suivante contient un résumé des principaux éléments qui caractérisent une bonne évaluation d'impact :

- ✓ Une question de politique concrète (fondée sur une théorie du changement) à laquelle l'évaluation d'impact peut fournir une réponse
- ✓ Une stratégie d'identification (ou méthodologie d'évaluation) valide, compatible avec les règles opérationnelles du programme, qui illustre la relation causale entre le programme et les résultats à l'étude
- ✓ Un échantillon avec une puissance suffisante pour détecter des impacts significatifs du point de vue politique et un échantillon représentatif qui permet de généraliser les résultats à une population plus étendue
- ✓ Une base de données de qualité fournissant les variables requises pour l'analyse, incluant à la fois des données de référence et des données de suivi, tant pour le groupe de traitement que pour le groupe de comparaison
- ✓ Une équipe d'évaluation bien organisée qui travaille en étroite collaboration avec les décideurs et gestionnaires du programme
- ✓ Un rapport d'impact et des notes de synthèse politique diffusés rapidement au public cible, qui fournissent des informations pertinentes pour la conception du programme et qui alimentent les dialogues de politique.

Nous soulignons ci-après quelques conseils formulés dans cet ouvrage pour limiter les risques auxquels les évaluations d'impact font souvent face :

- ✓ Il est largement préférable de concevoir l'évaluation d'impact au début du cycle de projet dans le cadre de la conception du programme. Une planification menée suffisamment tôt permet de concevoir une évaluation prospective fondée sur la meilleure méthodologie et laisse le temps nécessaire pour collecter des données de référence avant le lancement du programme dans les zones évaluées.
- ✓ Les résultats doivent être étayés par des données complémentaires provenant d'évaluations de processus et de données de suivi qui fournissent une image claire de la mise en œuvre du programme. Si un programme est efficace, il est important de comprendre pourquoi. Si un programme échoue, il est important de pouvoir distinguer entre un programme mal mis en œuvre et un programme mal conçu.
- ✓ Collectez des données de référence et intégrez une méthode de rechange à votre plan d'évaluation. Si la méthode d'évaluation initialement prévue n'est pas valide (par exemple si le groupe de comparaison initial bénéficie du programme), un plan de rechange peut éviter de devoir renoncer entièrement à l'évaluation.

- ✓ Conservez un identifiant unique pour chaque unité dans toutes les bases de données afin de pouvoir exploiter facilement toutes les ressources disponibles au moment de l'analyse. Par exemple, un ménage donné doit avoir le même identifiant tant dans les systèmes de suivi que dans les enquêtes de référence et de suivi.
- ✓ Les évaluations d'impact sont utiles aussi bien pour comprendre comment un programme fonctionne et éprouver différentes alternatives de conception de programmes que pour évaluer l'impact global d'un programme au bénéfice unique. La désagrégation des divers éléments d'un programme, même universel et très étendu, peut être un excellent moyen d'apprendre et de tester des innovations dans le cadre d'évaluations d'impact bien conçues. Le développement d'une innovation en tant que projet pilote à petite échelle dans le contexte d'une évaluation plus étendue peut fournir de précieuses informations pour les prises de décision futures.
- ✓ Les évaluations d'impact doivent être pleinement considérées comme l'une des composantes du programme ; il faut y consacrer le personnel et le budget adéquats ainsi que des ressources techniques et financières suffisantes. Soyez réaliste quant aux coûts et à la complexité que représente une évaluation d'impact. La conception de l'évaluation et la collecte des données de référence peuvent durer environ un an. Une fois le programme lancé, il faut une période d'exposition suffisante avant que l'intervention n'affecte les résultats. Selon le programme, cette période peut s'étendre entre un à cinq ans, voire plus. La collecte d'une ou de plusieurs enquêtes de suivi, la réalisation des analyses et la diffusion des résultats nécessitent également des efforts importants sur plusieurs mois. Un cycle d'évaluation d'impact complet nécessite généralement au moins trois ou quatre ans d'efforts soutenus. Des ressources financières et techniques adéquates sont nécessaires à chaque étape du processus.

Au final, les évaluations d'impact fournissent des réponses concrètes à des questions de politique spécifiques. Même si les réponses sont taillées en fonction des besoins de l'entité qui commandite et finance l'évaluation, d'autres organismes à travers le monde peuvent en tirer des enseignements et les utiliser dans leurs propres prises de décisions. Par exemple, plusieurs récents programmes de transferts monétaires conditionnels en Afrique, en Asie et en Europe ont tiré des enseignements des évaluations novatrices des programmes Familias en Acción (Colombie), Progresa (Mexique) et d'autres programmes de transferts monétaires mis en œuvre en Amérique latine. Dans ce sens, les évaluations d'impact peuvent être considérées comme un bien public global. Les conclusions d'une évaluation alimentent les connaissances globales sur le sujet en question. Cet ensemble de preuves peut par la suite être utilisé par d'autres pays et dans d'autres contextes pour formuler des décisions de politique éclairées. Dans ce sens, la communauté internationale renforce de plus en plus son support aux initiatives d'évaluations rigoureuses.

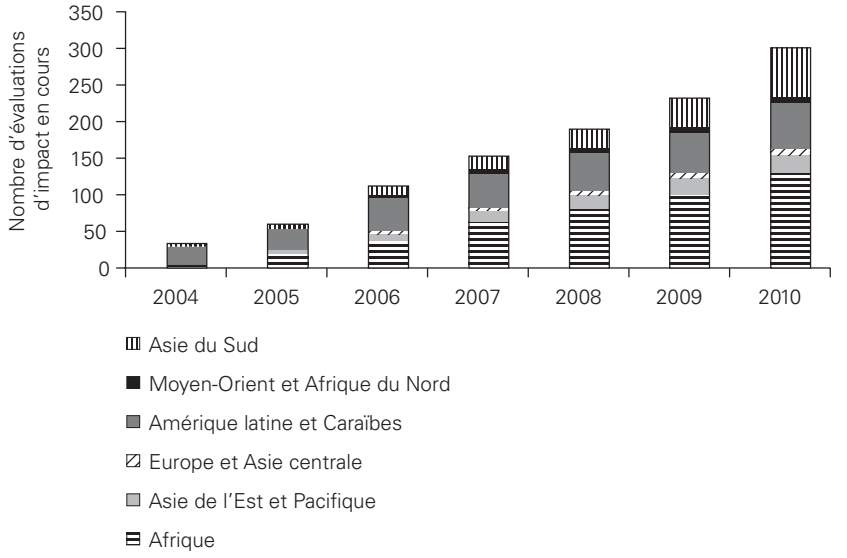
À l'échelle nationale, les gouvernements les plus avertis et exigeants cherchent à démontrer à leurs citoyens les résultats obtenus et à rendre des comptes des performances de leurs politiques. Il est de plus en plus fréquent que des évaluations soient réalisées par des ministères nationaux ou des entités locales spécialement créés pour coordonner un programme d'évaluation national, à l'image du Conseil national d'évaluation de la politique de développement social (CONEVAL) au Mexique et du Département de suivi et d'évaluation de la performance en Afrique du Sud. De plus en plus, les conclusions et les preuves générées par les évaluations d'impact sont prises en compte pour informer les décisions budgétaires des congrès nationaux. Dans les systèmes où les programmes sont évalués sur la base de preuves tangibles en fonction de leurs impacts sur des résultats finaux, les programmes qui génèrent des données positives pourront être soutenus tandis que ceux qui produisent peu d'informations sur leur efficacité auront du mal à trouver des financements.

Les institutions multilatérales comme la Banque mondiale et les banques de développement régionales ainsi que les agences de développement nationales, les États donateurs et les organismes philanthropiques exigent eux aussi des preuves plus nombreuses et plus concrètes sur l'efficacité des ressources de développement. Ces preuves sont un moyen de rendre compte aux organismes prêteurs ou donateurs de la performance des politiques mises en œuvre et d'orienter les prises de décisions concernant l'allocation des ressources de développement. Le nombre d'évaluations d'impact réalisées par les institutions de développement a fortement augmenté au cours des dernières années. La figure 14.1 indique le nombre d'évaluations d'impact en cours ou effectuées par la Banque mondiale entre 2004 et 2010, par région. Cette tendance positive devrait se maintenir.

Un nombre croissant d'institutions spécialisées dans la réalisation d'évaluations d'impact de qualité prospèrent, notamment dans la sphère universitaire, à l'image de Poverty Action Lab, d'Innovations for Poverty Action, du Center of Evaluation for Global Action ou des organismes indépendants qui soutiennent les évaluations d'impact comme l'International Initiative for Impact Evaluation. Plusieurs associations d'évaluation d'impact regroupent des spécialistes, des chercheurs et des décideurs intéressés par ce thème, parmi lesquelles le Network of Networks on Impact Evaluation et des associations régionales comme l'African Evaluation Association et le Réseau d'évaluation d'impact de la Latin American and Caribbean Economics Association. Tous ces efforts illustrent l'importance croissante de l'évaluation d'impact dans la politique de développement international[1].

Par conséquent, que vous soyez un professionnel de l'évaluation d'impact, que vous commanditiez des évaluations d'impact ou que vous exploitiez leurs résultats pour vos prises de décision, il est aujourd'hui indispensable pour tout spécialiste du développement de comprendre le langage qui y est associé. Les preuves rigoureuses générées par les évaluations d'impact constituent un catalyseur du dialogue sur les politiques de développement et contribuent à justifier le bien-fondé des décisions d'investissement dans des programmes et des politiques de développement.

Figure 14.1 Nombre d'évaluations d'impact effectuées par la Banque mondiale par région, 2004 2010



Source : Banque mondiale.

Les conclusions des évaluations d'impact permettent aux gestionnaires de projet de prendre des décisions éclairées sur la façon d'atteindre les résultats visés de la manière la plus rentable. Forts de ces conclusions, les décideurs peuvent boucler la boucle en intégrant les résultats des évaluations au processus de prise de décision. Ce type de preuves peut mieux informer les débats, les opinions et, au final, les décisions d'allocation de ressources humaines et monétaires prises par les gouvernements, les institutions multilatérales et les donateurs.

L'élaboration de politiques fondées sur des preuves consiste essentiellement à reprogrammer les budgets pour étendre les programmes rentables, réduire les programmes inefficaces et améliorer la conception des programmes en se fondant sur les meilleures données disponibles. En ce sens, l'évaluation d'impact n'est pas un exercice purement théorique. Elle répond au besoin de trouver des réponses à des questions de politique affectant la vie quotidienne des populations. Les décisions sur la manière optimale d'allouer des ressources limitées à des programmes de lutte contre la pauvreté, de santé, d'éducation, de sécurité sociale, de microcrédit, de développement agricole, etc. ont le potentiel d'améliorer le bien-être des populations à travers le monde. Il est essentiel que ces décisions soient fondées sur les informations et les preuves les plus rigoureuses possible.

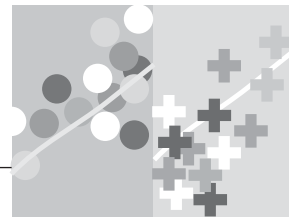
Note

1. Pour en savoir plus, voir Savedoff, Levine et Birdsall (2006).

Références

Legovini, Arianna. 2010. « Development Impact Evaluation Initiative: A World Bank–Wide Strategic Approach to Enhance Development Effectiveness. » Rapport préliminaire aux Vice-présidents, Opérations, Banque mondiale, Washington, DC.

Savedoff, William, Ruth Levine et Nancy Birdsall. 2006. « When Will We Ever Learn? Improving Lives through Impact Evaluation. » CGD Evaluation Gap Working Group Paper, Center for Global Development, Washington, DC. http://www.cgdev.org/content/publications/detail/7973.



GLOSSAIRE

Les termes en italique sont définis dans le présent glossaire.

Activité. Actions prises ou travaux réalisés à travers lesquels des *intrants*, comme des fonds, de l'assistance technique ou d'autres types de ressources, sont mobilisés pour produire des *extrants*.

Analyse coût bénéfice (ou analyse coût avantage). Calcul ex ante des coûts et des bénéfices espérés, servant à évaluer des propositions de projets. Dans le cadre d'une *évaluation d'impact*, on peut calculer les coûts et bénéfices ex post si les bénéfices sont quantifiables en termes monétaires et que des données sur les coûts sont disponibles.

Appariement (« matching »). L'appariement est une méthode d'*évaluation* non expérimentale où l'on constitue le meilleur *groupe de comparaison* possible pour un *groupe de traitement* donné à l'aide de grandes bases de données et de techniques statistiques complexes.

Attrition. Une attrition se produit lorsqu'il y a une déperdition de certaines unités de *l'échantillon* d'une ronde à l'autre de la collecte des données ; par exemple si les migrants ne sont pas suivis. L'attrition est un cas de *non-réponse* totale ou unitaire. L'attrition peut causer un *biais* dans les *évaluations d'impact* lorsqu'elle est corrélée avec le traitement.

Biais. Le biais d'un *estimateur* est la différence entre la valeur espérée du paramètre estimé et la valeur réelle de ce dernier. Dans le cadre d'une *évaluation d'impact*, il s'agit de la différence entre l'impact calculé et l'impact réel du programme.

Biais de sélection. Le biais de sélection se produit lorsque les raisons pour lesquelles un individu participe au programme sont corrélées aux résultats. Ce biais se produit souvent lorsque le *groupe de comparaison* est constitué d'individus qui ne sont pas éligibles pour participer au programme ou qui choisissent volontairement de ne pas y participer.

Cadre d'échantillonnage (ou base d'échantillonnage). La liste la plus exhaustive qu'on puisse obtenir des unités constituant une *population à l'étude*. Toute différence entre le cadre d'échantillonnage et la population à l'étude donne lieu à un *biais d'échantillonnage* (biais de couverture). Si un *biais* de couverture existe, les résultats obtenus à partir de *l'échantillon* n'ont pas de *validité externe* pour l'ensemble de la population à l'étude.

Calculs de puissance. Les calculs de *puissance* indiquent la taille que doit avoir *l'échantillon* pour détecter *l'effet minimal désiré* dans une *évaluation*. Les calculs de puissance dépendent de paramètres comme la puissance (ou la probabilité d'une *erreur de type II*), le *seuil de signification*, la variance et la *corrélation intra-grappe* du *résultat* à l'étude.

Chaîne de résultats. Une chaîne de résultats décrit la logique de réalisation des objectifs de développement d'un programme. Elle montre les liens entre les *intrants* et les résultats en passant par les activités et les *extrants*.

Comparaison avant-après. Également appelée « comparaison pré-post » et « comparaison réflexive », la comparaison avant-après vise à évaluer l'impact d'un programme en procédant à un suivi de l'évolution des *résultats* obtenus par les participants au programme au fil du temps, en particulier en comparant les résultats avant et après sa mise en œuvre.

Contrefactuel. Le contrefactuel est une estimation de ce qu'aurait été le *résultat* (*Y*) pour un participant au programme en l'absence du programme (*P*). Par définition, le contrefactuel n'est pas observable. Il faut donc l'estimer en recourant à des *groupes de comparaison*.

Corrélation intra-grappe. La corrélation intra-grappe est la corrélation (ou l'association) des *résultats* ou des caractéristiques entre les unités d'une même grappe. Par exemple, les enfants qui fréquentent la même école proviennent d'ordinaire de la même zone d'habitation ou du même milieu socioéconomique, ce qui implique une source de corrélation.

Données d'enquête. Données qui correspondent à un *échantillon* de la population à l'étude. Se différencie des *données de recensement*.

Données de recensement. Données qui recouvrent toutes les unités de la population à l'étude. Se différencie des *données d'enquête*.

Double Différence. Également appelée « différence des différences » ou « DD ». La double différence estime le *contrefactuel* pour le changement du *résultat* dans le *groupe de traitement* par le changement du résultat dans le *groupe de comparaison*. Cette méthode permet de prendre en compte toute différence entre le groupe de traitement et le groupe de comparaison qui est invariable dans le temps. Les deux différences sont donc celle de l'avant et de l'après, et celle entre le groupe de traitement et le groupe de comparaison.

Échantillon. En statistique, un échantillon est un sous-ensemble d'une population. En règle générale, la taille de la population est très grande, ce qui rend son *recensement*, c'est-à-dire une énumération exhaustive de toutes ses unités, impraticable ou impossible. Les chercheurs prélèvent à la place à l'aide d'un cadre d'échantillonnage un sous-ensemble représentatif de la population, et recueillent des statistiques pour cet échantillon. Ces statistiques permettent alors d'inférer ou d'extrapoler les paramètres pour l'ensemble de la population. Le processus par lequel l'échantillon est obtenu de la population à l'étude s'appelle *l'échantillonnage*.

Échantillon aléatoire. La meilleure façon d'éviter qu'un *échantillon* soit biaisé ou non représentatif est de le prélever de façon aléatoire. Un échantillon aléatoire est un échantillon probabiliste pour lequel toutes les unités de la population à l'étude ont la même probabilité d'être sélectionnées.

Échantillon par grappes. *Échantillon* obtenu par le prélèvement d'un *échantillon aléatoire* de grappes, après quoi soit l'ensemble des unités des grappes sélectionnées constitue l'échantillon, soit un certain nombre d'unités est sélectionné de manière aléatoire dans chaque *grappe* prélevée. Chaque grappe a une probabilité bien définie d'être sélectionnée, et les unités sélectionnées de chaque grappe ont elles aussi une probabilité bien définie d'être prélevées.

Échantillon stratifié. Échantillon obtenu en répartissant la population à l'étude (*cadre d'échantillonnage*) en strates ou groupes (p. ex. groupe d'hommes et groupe de femmes), et en prélevant ensuite un *échantillon aléatoire* pour chaque groupe. Un *échantillon* stratifié est un échantillon probabiliste, c'est-à-dire que toutes les unités d'un même groupe (ou strate) ont la même probabilité d'être prélevées.

Échantillonnage. Processus par lequel des unités sont prélevées du *cadre d'échantillonnage* obtenu pour la *population à l'étude* (univers). Il existe plusieurs procédures d'échantillonnage. Les méthodes d'échantillonnage probabilistes sont les plus rigoureuses, car elles attribuent à chaque unité une probabilité bien définie d'être prélevée. L'échantillonnage aléatoire, l'échantillonnage aléatoire stratifié et l'échantillonnage par grappes sont toutes des méthodes d'échantillonnage probabiliste. Les échantillonnages non probabilistes, comme l'échantillonnage par jugement et l'échantillonnage de convenance, peuvent mener à des erreurs d'échantillonnage.

Effet. Changement intentionnel ou non dû directement ou indirectement à une intervention.

Effet Hawthorne. L'effet Hawthorne se produit lorsque les unités changent de comportement du fait même d'être observées.

Effet John Henry. L'*effet* John Henry se produit lorsque les unités de comparaison font des efforts supplémentaires pour compenser l'absence du traitement. Lorsque l'on compare les unités de traitement aux unités de comparaison qui font des efforts supplémentaires, l'impact estimé du programme est biaisé ; c'est-à-dire que l'impact estimé est moindre que celui qu'on observerait si les unités de comparaison n'avaient pas fait d'effort supplémentaire.

Effet minimal désiré. Le changement minimal des *résultats* qui justifierait l'investissement consenti dans une intervention, prenant en compte non seulement le coût du programme et ses bénéfices, mais aussi son coût d'opportunité (les fonds n'ayant pas été investis ailleurs). L'*effet* minimal désiré est un paramètre qui entre dans les *calculs de puissance* : les *échantillons d'évaluation* doivent être de taille suffisante pour permettre de détecter l'effet minimal désiré à une certaine *puissance*.

Effets de diffusion (ou effets de débordements ou « spillover »). Également appelé contamination s'ils affectent le *groupe de comparaison*. L'effet de diffusion se produit lorsque le groupe de comparaison est affecté par le traitement administré au *groupe de traitement*, même si le traitement n'est pas directement administré au groupe de comparaison. Si l'effet de diffusion sur le groupe de comparaison est négatif, c'est-à-dire si le programme nuit à ce groupe, la différence directe entre les *résultats* du groupe de traitement et ceux du groupe de comparaison produit une surestimation de l'impact du programme. Par contre, si l'effet de diffusion sur le groupe de comparaison est positif, c'est-à-dire si le programme est bénéfique pour ce groupe, le résultat est alors une sous-estimation de l'impact du programme.

Enquête de suivi. Également appelée « enquête post-intervention » ou « enquête ex post ». Enquête qui est réalisée après le démarrage du programme, après que les participants ont bénéficié du programme. Une *évaluation d'impact* peut comprendre plusieurs enquêtes de suivi.

Erreur de type I. Erreur commise en rejetant *l'hypothèse nulle* alors qu'elle est valable. Dans le contexte d'une *évaluation d'impact*, une erreur de type I est commise lorsqu'une *évaluation* conclut qu'un programme a un impact, c'est-à-dire que l'hypothèse nulle selon laquelle il n'y a aucun impact est rejetée, alors que le programme n'a, en réalité, aucun impact, et que l'hypothèse nulle est donc valable. Le *seuil de signification* détermine la probabilité de commettre une erreur de type I.

Erreur de type II. Erreur commise en acceptant (en ne rejetant pas) *l'hypothèse nulle* alors que celle-ci n'est pas valable. Dans le contexte d'une *évaluation d'impact*, une erreur de type II est commise lorsqu'une évaluation conclut qu'un programme n'a aucun impact, c'est-à-dire que l'hypothèse nulle selon laquelle il n'y a aucun impact n'est pas rejetée, alors que le programme a, en réalité, un impact, et que l'hypothèse nulle n'est donc pas valable. La probabilité de commettre une erreur de type II est égale à 1 moins la *puissance*.

Estimateur. Un estimateur est une statistique (une fonction des données observée *d'un échantillon* observables) qui sert à estimer un paramètre inconnu de la population. L'estimation est le résultat de l'application de la fonction à un échantillon de données.

Estimateur de l'intention de traiter ou de l'IDT. *L'estimateur* de l'IDT est la simple différence entre *l'indicateur de résultat Y* pour le groupe auquel on a offert le traitement et le même indicateur pour le groupe auquel on n'a pas offert le traitement. Se différencie de l'effet du *traitement sur les traités*.

Évaluation. Les évaluations sont des appréciations périodiques et objectives de projets ou de programmes ou de politiques prévus, en cours de réalisation ou réalisés. Les évaluations fournissent des informations sur des questions précises, souvent liées à la conception, à la mise en œuvre et aux résultats.

Évaluation d'impact. Une *évaluation* d'impact est une évaluation qui tente d'établir un lien causal entre un programme et des indicateurs de *résultats*. Une évaluation d'impact tente de savoir si le programme est directement responsable de changements dans les indicateurs de résultats à l'étude. Se différencie de *l'évaluation de processus*.

Évaluation de processus. Une *évaluation* de processus tente de déterminer la qualité ou le degré de performance des processus d'un programme, comme l'adéquation des procédures administratives, l'acceptabilité des bénéfices d'un programme, la clarté d'une campagne d'information, les mécanismes internes des organismes de mise en œuvre, leurs moyens d'action, leurs dispositifs de prestation de service, leurs pratiques de gestion. Se différencie de *l'évaluation d'impact*.

Extrant. Les biens ou services qui sont produits ou offerts directement par une intervention. Les extrants comprennent parfois des changements découlant de l'intervention et qui contribuent à l'obtention des *résultats*.

Grappe. Une grappe est un groupe d'unités qui se ressemblent d'une façon ou d'une autre. Dans un *échantillonnage* d'écoliers, par exemple, les enfants qui se rendent à la même école appartiennent à une même grappe car ils fréquentent les mêmes installations scolaires, ils disposent des mêmes enseignements et ils habitent le même quartier.

Groupe de comparaison. Également appelé « groupe de contrôle » ou « groupe témoin » dans le cadre d'un essai contrôlé randomisé. Un groupe de comparaison valable a les mêmes caractéristiques que le groupe de participants au programme (*groupe de traitement*), à la seule exception que les unités du groupe de comparaison ne participent pas au programme. Les groupes de comparaison servent à estimer le *contrefactuel*.

Groupe de traitement. Également appelé groupe d'intervention. Le groupe de traitement est le groupe des unités qui bénéficient d'une intervention, tandis que le *groupe de comparaison* n'en bénéficie pas.

Hypothèse. Une hypothèse est une explication avancée d'un phénomène observable. Voir également *hypothèse nulle* et *hypothèse alternative*.

Hypothèse alternative. Dans une *évaluation d'impact*, l'hypothèse alternative suppose généralement que l'*hypothèse nulle* est fausse, c'est-à-dire que l'intervention a un impact sur les *résultats*.

Hypothèse nulle. Une *hypothèse* nulle est une hypothèse falsifiable en utilisant des données observables. L'hypothèse nulle postule généralement une position par défaut. Dans le cadre d'une *évaluation d'impact*, la position par défaut est généralement qu'il n'y a aucune différence entre le groupe de traitement et le groupe de comparaison ou, en d'autres termes, que l'intervention n'a aucun impact sur les *résultats*.

Indicateur. Un indicateur est une *variable* qui mesure un phénomène à l'étude. Le phénomène peut être un *intrant*, un *extrant*, un *résultat*, une caractéristique ou un attribut.

Intrants. Les ressources financières, humaines et matérielles utilisées par une intervention ou un programme.

Ligne de base (ou enquête de référence). Pré-intervention, ex ante. La situation qui prévaut avant l'intervention, par rapport à laquelle l'évolution est mesurée et les comparaisons sont faites. La ligne de base (ou enquête de référence) est collectée avant la mise en œuvre du programme ou de la politique à évaluer afin d'obtenir une mesure des résultats en amont.

Méthodes de sélection aléatoire. Les « méthodes de sélection aléatoire » désignent un ensemble de méthodes où la sélection aléatoire est employée pour estimer le *contrefactuel*. On compte notamment parmi ces méthodes l'*assignation aléatoire* du traitement, l'*offre aléatoire* du traitement et la *promotion aléatoire*.

Modèle de discontinuité de la régression. Le modèle de discontinuité de la régression est une méthode d'*évaluation* non expérimentale. Elle convient aux programmes qui utilisent un indice continu pour classifier les participants potentiels et un seuil bien défini pour identifier les bénéficiaires. Le seuil d'éligibilité au programme est un seuil qui sépare le *groupe de traitement* et le *groupe de comparaison*.

Non-réponse. L'absence ou le manque de données pour certaines unités d'un *échantillon* constituent une non-réponse. La non-réponse unitaire se produit lorsqu'on ne possède aucune information pour certaines unités de l'échantillon ; c'est-à-dire quand l'échantillon prélevé est différent de l'échantillon prévu. L'*attrition* est une forme de non-réponse unitaire ou totale (« unit non-response »). La non-réponse partielle (« item non-response ») se produit lorsque les données sont incomplètes pour certaines unités prélevées. La non-réponse peut créer un *biais* dans les résultats de l'*évaluation* si elle est corrélée avec le traitement.

Offre aléatoire. L'offre aléatoire est une méthode qui permet de déterminer l'impact d'une intervention. L'intervention est offerte aux personnes éligibles de manière aléatoire de façon à ce qu'elles aient toutes la même chance de participer au programme. Même si l'administrateur du programme peut sélectionner au hasard, parmi toutes les unités éligibles, celles à qui offrir le traitement, il ne peut parfois pas obtenir une conformité absolue. Il ne peut pas forcer une unité à participer ou à accepter, ni refuser la participation à une unité qui insiste pour participer. Dans ce contexte, l'offre aléatoire du programme sert de *variable instrumentale* pour la participation réelle au programme.

Population à l'étude. Le groupe d'unités qui est éligible pour recevoir l'intervention ou du traitement. La population à l'étude est parfois appelée « univers ».

Promotion aléatoire. La promotion aléatoire est une méthode proche de celle de l'*offre aléatoire*. Au lieu de sélectionner au hasard les unités auxquelles on offre le traitement, les unités sont sélectionnées au hasard pour recevoir une promotion et ainsi augmenter la probabilité qu'elles participent au traitement. De cette façon, le programme demeure ouvert à tous.

Puissance. La puissance est la probabilité d'observer un impact s'il existe. La puissance d'un test est égale à un moins la probabilité d'une *erreur de type II*, allant de zéro à un. La puissance varie le plus souvent entre 0,8 et 0,9. Les valeurs élevées de la puissance sont plus conservatrices. Elles réduisent le risque des erreurs de type II. La puissance d'une *évaluation d'impact* est élevée si le risque de ne pas observer d'impacts, c'est-à-dire de commettre une erreur de type II, est faible.

Puissance statistique. La *puissance* d'un test statistique est la probabilité que le test aboutisse au rejet de l'*hypothèse nulle* lorsque l'*hypothèse alternative* est valable (c'est-à-dire qu'aucune *erreur de type II* n'est commise). Le risque de commettre une erreur de type II décroit au fur et à mesure que la puissance augmente. La probabilité de commettre une erreur de type II est désignée par le taux de faux-négatif (β). La puissance est donc égale à 1 – β.

Rapport coût-efficacité. Pour déterminer le rapport coût-efficacité, il faut comparer des interventions similaires sur les plans du coût et de l'efficacité. Ainsi, les *évaluations d'impact* de divers programmes éducatifs permettent aux décideurs de prendre des décisions éclairées sur l'intervention qui permet de produire les résultats souhaités au moindre coût et en fonction des contraintes qui sont les leurs.

Régression. En statistique, l'analyse de régression comprend l'ensemble des techniques pour modéliser et analyser plusieurs *variables* en considérant le lien entre une variable dépendante et une ou plusieurs variables indépendantes. Dans l'*évaluation d'impact*, l'analyse de régression permet de comprendre comment l'*indicateur de résultat Y* (variable dépendante) évolue en fonction de l'affectation au traitement, ou *groupe de comparaison P*, (variable indépendante) alors que les caractéristiques des participants (variables indépendantes) ne changent pas.

Résultat. Intermédiaire ou final. Un résultat est le produit de l'interaction entre des facteurs d'offre et de demande. Par exemple, si une intervention renforce l'offre des services de vaccination, le nombre de vaccinations constitue alors un résultat, celui-ci ne dépendant pas seulement de l'offre en vaccins, mais aussi du comportement des personnes ciblées : se rendent-elles au centre de vaccination pour se faire vacciner ? Les résultats finaux et les résultats à long terme sont plus distants, soit dans la dimension temporelle (une longue période est nécessaire pour arriver au résultat), soit dans la dimension causale (un grand nombre de liens de cause à effet sont nécessaires pour atteindre le résultat).

Sélection aléatoire (ou essai contrôlé randomisé). La sélection aléatoire est considérée comme la méthode la plus rigoureuse pour estimer le *contrefactuel*. Elle est souvent décrite comme « l'étalon-or » de *l'évaluation d'impact*. Les bénéficiaires de l'intervention sont sélectionnés au hasard parmi la population éligible. Tous les individus éligibles ont donc la même chance de participer au programme. Avec des *échantillons* de taille suffisante, la sélection aléatoire garantit que les caractéristiques, observées et non observées des groupes de traitement et de contrôle soient semblables, éliminant ainsi le *biais de sélection*.

Seuil de signification. Le seuil de signification est généralement désigné par la lettre grecque α (alpha). Les seuils de signification les plus courants sont 5 % (0,05), 1 % (0,01) et 0,1 % (0,001). Si un test de signification produit une valeur *p* inférieure au seuil α, l'*hypothèse nulle* est rejetée. Un tel résultat est qualifié de manière informelle comme étant « statistiquement significatif ». Plus le seuil de signification est petit, plus la preuve requise doit être forte. Le choix du seuil de signification est arbitraire. Mais le seuil de 5 % est conventionnel.

Suivi. Le suivi est un processus continu de collecte et d'analyse d'informations dans le but de déterminer la performance du projet, du programme ou de la politique mis en œuvre. Ce processus s'appuie essentiellement sur les données administratives pour comparer la performance effective aux résultats espérés, les programmes entre eux et pour analyser leurs tendances dans le temps. Le suivi se concentre généralement sur les *intrants*, les activités et les *extrants*, ainsi qu'occasionnellement les *résultats*. Le suivi est utile pour la gestion quotidienne du programme.

Traitement sur les traités (effet du). Également appelé *estimateur* TT. L'*effet* du traitement sur les traités désigne l'impact du traitement sur les unités qui ont effectivement reçu le traitement. Se différencie de l'*intention de traiter*.

Validité externe. L'estimation de l'impact causal du programme a une validité externe si elle est généralisable à l'univers de toutes les unités éligibles. Pour qu'une évaluation ait une validité externe, l'*échantillon* de l'évaluation doit être représentatif de l'univers des unités éligibles.

Validité interne. Une *évaluation d'impact* a une validité interne si elle se fonde sur un *groupe de comparaison* valide, c'est-à-dire un groupe de contrôle qui fournit une estimation valide du *contrefactuel*.

Variable. Dans la terminologie statistique, une variable est un symbole qui représente une valeur changeante.

Variable instrumentale. Une variable instrumentale est une *variable* qui permet d'estimer l'impact causal d'un programme lorsque la participation au programme est déterminée en partie par les participants potentiels. Pour être considérée comme une variable instrumentale valable, une variable doit posséder deux caractéristiques : 1) elle doit être corrélée avec la participation au programme, et 2) elle ne doit pas être corrélée avec les *résultats Y* (sauf à travers la participation au programme), ni avec les variables non observées.

ECO-CONTRÔLE
Déclaration d'avantages environnementaux

La Banque Mondiale a pris l'engagement de préserver les forêts et les ressources naturelles. La maison d'édition a décidé d'imprimer L'evaluation d'impact en practique sur du papier recyclé comprenant 50 pourcent de papier déjà utilisé, selon les standards recommandés par Green Press Initiative, un programme à but non lucratif incitant les maisons d'édition à utiliser du bois qui ne provienne pas de forêts en danger. Pour plus d'informations, vous pouvez visiter www.greenpressinitiative.org.

Sauvés:

- **8 arbres**
- **2 millions BTU**
- **327 kg d'effet de serre net**
- **13.128 litres d'eau usée**
- **96 kg de déchets solides**

« Cet ouvrage constitue un guide pratique, complet et clair sur l'évaluation d'impact. Son contenu, qui traite des raisons de procéder à des évaluations d'impact, des avantages des différentes méthodologies, en passant par les calculs de puissance et les coûts, est présenté de manière très claire et couvre un grand nombre de domaines. Ce manuel deviendra un guide de référence incontournable et influencera l'élaboration des politiques pour les années à venir. »

Orazio Attanasio, *Professor of Economics, University College London; Director, Centre for the Evaluation of Development Policies, Institute for Fiscal Studies, Royaume-Uni.*

« Ce précieux ouvrage s'adresse à celles et ceux qui visent à mener des évaluations d'impact dans les pays en développement. Il décrit les enjeux conceptuels et pratiques des évaluations en s'appuyant sur des exemples tirés d'expériences récentes. »

Michael Kremer, *Gates Professor of Developing Societies, Department of Economics, Harvard University, États-Unis.*

« Les ingrédients de base indispensables à la réussite des évaluations de politiques publiques sont a) des méthodologies appropriées, b) la capacité à résoudre des problèmes pratiques tels que la collecte de données, les limites budgétaires ou la rédaction du rapport final et c) la responsabilisation des gouvernements. Cet ouvrage présente des outils méthodologiques solides pour évaluer l'impact des programmes publics. Il expose aussi de nombreux exemples et nous emmène au cœur de la mise en œuvre des évaluations d'impact, de l'étape qui consiste à convaincre les décideurs à celle de la diffusion des résultats. Si davantage de praticiens et de décideurs lisent ce manuel, nous aurons de meilleures politiques et de meilleurs résultats dans de nombreux pays. Si les gouvernements se responsabilisent aussi davantage, l'impact de ce manuel n'en sera que plus important. »

Gonzalo Hernández Licona, *Executive Secretary, National Council for the Evaluation of Social Development Policy (CONEVAL), Mexique.*

« Je recommande cet ouvrage comme un guide clair et accessible pour faire face aux défis pratiques et techniques inhérents à la conception des évaluations d'impact. Le manuel est fondé sur des ressources éprouvées lors d'ateliers conduits à travers le monde et constitue une référence utile tant pour les praticiens, que pour les décideurs ou les évaluateurs. »

Nick York, *Head of the Evaluation Department, Department for International Development, Royaume-Uni.*

« La connaissance est un atout essentiel pour comprendre la nature complexe du processus de développement. Les évaluations d'impact contribuent à combler le fossé entre l'intuition et les preuves et ainsi à améliorer l'élaboration de politiques publiques. Cet ouvrage est l'un des produits concrets du Fonds espagnol pour l'évaluation d'impact. Il munit les praticiens en matière de développement humain d'outils de pointe qui leur permettront de générer des preuves au sujet de quelles politiques sont efficaces et pourquoi. Parce qu'il améliore notre capacité à atteindre des résultats, cet ouvrage devrait transformer en profondeur les pratiques de développement. »

Soraya Rodríguez Ramos, *Secretary of State for International Cooperation, Espagne.*

BANQUE MONDIALE

SIEF Fonds espagnol d'évaluation d'impact

ISBN 978-0-8213-8752-8

SKU 18752