DOCUMENT DE RECHERCHE

EPEE

CENTRE D'ETUDES DES POLITIQUES ECONOMIQUES DE L'UNIVERSITE D'EVRY

Evaluation aléatoire et expérimentations sociales

Yannick L'Horty & Pascale Petit

10 - 02

www.univ-evry.fr/EPEE

Université d'Evry Val d'Essonne, 4 bd. F. Mitterrand, 91025 Evry CEDEX

Evaluation aléatoire et expérimentations sociales

Yannick L'Horty [◊], Pascale Petit [↑]

Octobre 2010

Résumé [1]

Les méthodes d'évaluation aléatoire ont commencé à être appliquées à de grands programmes sociaux en France depuis 2007, soit plus de quarante ans après les premiers travaux analogues menés aux Etats-Unis. Elles connaissent depuis un développement rapide. Ces méthodes expérimentales consistent à évaluer les effets d'une politique en comparant un groupe test à un groupe témoin, constitués par tirage au sort. Dans cet article dont l'objet est de proposer une introduction à l'application de ces méthodes aux expérimentations sociales, nous soulignons que leurs spécificités ne se réduisent pas (i) au principe de l'assignation aléatoire. Elles ont aussi pour autres singularités qui les distinguent des autres méthodes d'évaluation : ii) la dimension expérimentale du programme évalué, iii) le partenariat particulier qui est organisé entre le chercheur et l'institution expérimentatrice, iv) la conception a priori *d'un protocole d'évaluation et d'un système d'observation et de traitement de l'information. Nous montrons que chacune de ces quatre singularités est la source d'un apport particulier de ces méthodes mais a aussi pour contrepartie une limite spécifique qui mérite d'être soulignée.*

Codes JEL : C93, I38.

Mots-clés : Expérimentation, évaluation de politiques publiques, randomisation, évaluation aléatoire

---

[◊] Université de Paris Est Marne la Vallée, ERUDITE, CEE et TEPP-CNRS. yannick.lhorty@cee-recherche.fr
[↑] Université d'Evry-Val d'Essonne, EPEE, CEE, et TEPP-CNRS. pascale.petit@univ-evry.fr

[1] Cet article est tiré d'un travail de synthèse réalisé à la demande du Ministère de la Jeunesse et des Solidarités Actives, dont une toute première version a été diffusée sous la forme d'un guide méthodologique rédigé à l'attention des porteurs de projet d'expérimentations sociales. Il a bénéficié de remarques et de discussions avec François Bourguignon, Emmanuel Duguet, Marc Gurgand, François Langot, François Legendre, Bénédicte Rouland, Marie-Odile Simon, Augustin Vicard,. Il a bénéficié également des remarques des participants au colloque « Méthodes d'Enquêtes et d'Evaluation pour la Sociologie et l'Economie » (Le Touquet, mars 2009), au séminaire de l'EPEE (Evry, juin 2009) et à l'école thématique du CNRS « Evaluation des Politiques Publiques » (Aussois, mars 2010).

Randomized Evaluation and Social Experiments

Abstract

Randomized evaluation of public policies began to be applied to major social programs in France since 2007, more than forty years after the first similar work in the United States. They have been experiencing rapid development. These experimental methods are to evaluate the effects of a policy by comparing a test group with a control group, made by random assignment. In this paper whose purpose is to provide an introduction to the application of these methods to social experiments, we emphasize that their specificities are several and cannot be reduced to (i) the principle of random assignment. They also have other peculiarities that distinguish them from other evaluation methods: ii) the fact that the program is experimental, iii) the particular partnership between the researcher and the institution experimenter, iv) the fact that researchers have to built an explicit protocol and a system of observation and information processing before the experiment beginning. We show that each of these four singularities is a source of both methodological advantage and potential limitation that should be emphasized

JEL Codes : C93, I38.

Keywords: Social Experiment, public policies evaluation, randomization

Introduction

L'évaluation des politiques publiques est un thème qui paraît de prime abord assez structurant pour les sciences sociales dans la mesure où il permet de rendre apparente les différences d'approches à la fois entre disciplines et au sein des disciplines. Dans ce champ, il est fréquent de dire que les économistes sont plutôt spécialisés dans les méthodes quantitatives et statistiques, alors que les chercheurs d'autres disciplines des sciences sociales, tels les sociologues et les politistes, privilégient plus volontiers les approches qualitatives. Les premiers sont connus pour aimer exploiter de larges bases de données issues d'enquêtes statistiques ou de sources administratives, alors que les seconds sont réputés privilégier des investigations approfondies sur des petits échantillons raisonnés de l'ordre de quelques dizaines d'individus, à l'aide de techniques d'entretiens et d'expression des acteurs. Ces différences de méthodes ne renvoient pas uniquement à une différence dans la taille des échantillons mais aussi à des différences de questionnement et d'objet de recherche. Lorsqu'un économiste évalue les effets *ex post* d'une politique publique, il entend mesurer de façon précise et *ceteris paribus* son impact causal sur un certain nombre de variables d'intérêt ce qui implique de disposer d'un large ensemble d'observations. Les politistes ou les sociologues veulent plutôt évaluer des processus et des stratégies d'acteurs, ce qui requiert une observation approfondie et par conséquent un échantillon de taille réduite, de façon à pouvoir repérer les obstacles au bon fonctionnement d'une action publique et les leviers sur lesquels il importe d'agir pour en améliorer les effets. De ce point de vue, les deux types d'approches recouvrent des questionnements qui apparaissent finalement assez complémentaires.

L'une des particularités des méthodes d'évaluation aléatoire est de prendre à contre pied ce type de représentation, sans doute un peu manichéenne. Alors que ces méthodes sont surtout déployées par des économistes, elles peuvent s'appliquer à des échantillons d'assez petite taille, de l'ordre de quelques centaines d'individus, et même en l'absence de bases de données préalable. Elles sont d'ailleurs largement utilisées par l'économie du développement dans des contextes où l'on souhaite évaluer les effets d'un programme sans disposer de données d'enquête ou de sources administratives pré-existantes (pour un survol, voir Kremer [2003] et Duflo, Glennerster et Kremer [2006]). L'objet est d'étudier l'impact d'un programme local de développement, de façon précise et chiffrée, sans pouvoir mobiliser aucune donnée *a priori*. Pour y parvenir, il importe de créer *ex nihilo* un système d'observation et de recueil des données à des fins d'évaluation.

Ce n'est pas le moindre des paradoxes de ces méthodes qui occupent une position particulière dans le spectre des technologies d'évaluation des politiques publiques. Cette originalité de positionnement ajoute à leur intérêt essentiel qui est de pouvoir mesurer l'effet causal d'un programme à partir d'une base de données qui est dépourvue d'emblée de tous biais de sélection. Ce faisant, elles ouvrent des perspectives très attractives pour les évaluateurs. Alors que l'économétrie de l'évaluation a pour objet quasi-exclusif, dans les approches non expérimentales, le contrôle des biais de sélection, elle peut désormais s'en dispenser. Il ne s'agit pas de faire « l'économie de l'économétrie » mais d'utiliser l'économétrie de façon plus productive et pour autre chose que le contrôle des biais.

Les méthodes expérimentales sont utilisées depuis longtemps dans les sciences dures, en médecine, en agronomie ou même en marketing. Levitt et List (2008) dans leur survol historique, distinguent trois générations d'évaluation aléatoire d'expériences de terrain. La première remonte aux travaux de Neyman et Fisher dans les années 1920 et 1930 où l'évaluation aléatoire est pour la première fois conçue comme un outil permettant d'identifier des effets causals et est appliquée en agronomie. La deuxième génération est celle des expérimentations sociales de grande échelle à partir des années soixante, où l'objet de l'expérimentation n'est plus des terres agricoles mais des groupes de personnes. En référence aux premiers travaux agronomiques, on parle d'essais de terrain (« Field Trials ») pour désigner ces méthodes appliquées au social (Burtless, 1995). Plus récemment, un troisième âge de l'expérimentation aurait été ouvert avec un élargissement considérable de leurs domaines d'application (au développement, à l'éducation, à la lutte contre la pauvreté, à la santé…), et du nombre et des types de questions traitées.

Les exemples les plus cités d'évaluations aléatoires de grands programmes sociaux viennent tous d'Amérique du Nord : l'expérimentation du New Jersey menée en 1968 pour tester un dispositif d'impôt négatif, suivie de trois autres expérimentations aux Etats-Unis au début des années soixante-dix ; le *Self Sufficiency Project* qui est une prime donnée à des bénéficiaires d'aide sociales pour les inciter au retour à l'emploi, expérimentée dans deux provinces canadiennes à partir de 1994 (Nouveau Brunswick et Colombie britannique) ; le programme *Moving to Opportunity*, mis en œuvre entre 1994 et 1998 pour favoriser la mobilité résidentielle des ménages pauvres dans cinq villes des Etats-Unis (Baltimore, Boston, Chicago, Los Angeles et New York) ; le *Progresa-Oportunidades* qui encourage depuis 1997

la scolarisation des enfants pauvres au Mexique. Ces méthodes sont désormais mises en œuvre dans tous les pays du nord de l'Europe, en Australie et dans de nombreux pays en développement, pour évaluer des programmes dans des domaines très variés (accès à l'emploi, lutte contre la pauvreté, amélioration des pratiques sanitaires, etc.).

Une abondante littérature leur est consacrée avec des publications dans les meilleures revues généralistes (pour une synthèse sur les programmes d'aides à l'emploi, voire Fougère, 2000). Ce champ de recherche fécond est aussi un champ en débat avec une controverse sur les portées et limites de ces méthodes qui opposent les défenseurs pragmatiques des *Randomized Field Trials,* des *Randomized Studies* et de la *New Development Economics* (Abhijit Banerjee, Gary Burtless, Michael Kremer, Esther Duflo,…) aux tenants des approches structurelles de l'évaluation des politiques publiques qui accordent une plus large place aux *a priori* de la théorie et au formalisme, notamment Angus Deaton (2009) ou Dani Rodrik (2008).

Nous nous intéressons ici au champ des politiques sociales au sens large et au cas de la France en particulier où l'introduction des méthodes d'évaluation aléatoire est très récente mais où sa diffusion est néanmoins massive et rapide. La première évaluation aléatoire de grande taille réalisée en France porte sur une expérimentation qui a eu lieu à l'automne 2007. Elle a été mise en œuvre par le CREST, l'Ecole d'Economie de Paris et le *Jameel – Poverty Action Laboratory* pour évaluer les effets des opérateurs privés d'accompagnement des demandeurs d'emploi inscrits à l'ANPE (Behaghel, Crépon et Gurgand, 2009). Quarante années séparent ainsi les expériences françaises et américaines en matière d'évaluation aléatoire. Depuis cette première étude, des dizaines de programmes sociaux innovants font l'objet d'évaluations aléatoires chaque année.

Pour expliquer la rapidité de cette expansion, il faut sans doute évoquer la conjonction d'un choc d'offre et d'un choc de demande, tous deux positifs. Coté offre, la diffusion des travaux d'Esther Duflo a vraisemblablement joué un rôle crucial en suscitant l'engouement des chercheurs français. L'un des principaux messages de la professeure au MIT qui est l'une des animatrices du réseau international du J-PAL est que, puisque l'évaluation expérimentale a fait ses preuves pour analyser les causes de la pauvreté dans les pays pauvres, il s'agit

maintenant de l'utiliser pour le même objectif dans les pays riches et notamment en France [2]. Côté demande, la commande publique a joué un rôle décisif, sous l'impulsion de Martin Hirsch. Devenu Haut Commissaire aux Solidarités Actives en juin 2007 et également Haut Commissaire à la Jeunesse en janvier 2009, l'initiateur du RSA va soutenir de façon constante le développement des expérimentations sociales[3] et de leur évaluation. Un premier appel à projet d'expérimentation sociale est lancé en 2007 avec un budget de 6 Millions €. Il est suivi en 2009 par une série d'appels à projets lancée par le fonds d'expérimentations pour la jeunesse (créé par l'article 25 de la loi généralisant le RSA du 1er décembre 2008) avec un budget total, issu d'un partenariat public-privé, de 150 millions d'Euros. Plus de 400 projets innovants sont ainsi financés qui prévoient fréquemment, mais pas systématiquement[4], une évaluation aléatoire.

L'évolution du cadre juridique et institutionnel a joué également un rôle important dans le développement des expérimentations sociales en France. Plusieurs obstacles législatifs et réglementaires ont du être levés pour rendre possible ce développement qui implique en pratique une rupture temporaire au principe d'égalité. Un cadre juridique est donné par la réforme constitutionnelle de décentralisation de 2003 et l'adoption la même année de la loi organique relative à l'expérimentation par les collectivités territoriales. Les expérimentations sociales deviennent possibles dès lors qu'elles ont un objet circonscrit et une durée limitée dans le temps et si elles sont menées en vue d'une généralisation. Elles doivent s'effectuer à l'initiative des collectivités locales et doivent nécessairement faire l'objet d'une évaluation. En pratique, l'expérimentation du revenu de solidarité active (RSA) prévue dans la loi du 21 août 2007 en faveur du travail, de l'emploi et du pouvoir d'achat (« loi Tepa ») va constituer la

---

[2] Voire la leçon inaugurale de la chaire internationale "Savoirs contre pauvreté » du collège de France (Duflo 2009) et les deux ouvrages Duflo 2010-a et 2010-b. Le message a été relayé par de nombreux économistes français, notamment François Bourguignon, le directeur de l'Ecole d'Economie de Paris qui préside le comité national d'évaluation des expérimentations du RSA, Marc Gurgand, qui préside le Conseil Scientifique du Fonds d'expérimentation pour la jeunesse, ou encore Bruno Crépon ou Philippe Zamora, du CREST. En outre, de nombreux colloques ont popularisé auprès d'un large public l'apport des méthodes expérimentales. On peut citer notamment les rencontres de l'insertion lancée à Grenoble en novembre 2007, le colloque « Expérimentations pour les politiques publiques de l'emploi et de la formation », organisé par la DARES en mai 2008, et la conférence nationale de l'expérimentation sociale, organisée par les deux Hauts Commissariats en mars 2010.

[3] Dans le cadre du Grenelle du l'insertion, une expérimentation sociale est définie comme « une innovation de politique sociale initiée dans un premier temps à petite échelle, compte tenu des incertitudes existantes sur ses effets et mise en œuvre dans des conditions qui permettent d'en évaluer les résultats, dans l'optique d'une généralisation si ces résultats s'avèrent probants ».

[4] Beaucoup de projets sont de très petite taille et les effectifs traités sont apparus insuffisants aux évaluateurs pour envisager une évaluation aléatoire.

première expérimentation sociale de grande ampleur en France, même si cette expérimentation n'a finalement pas été évaluée selon une méthode expérimentale [5].

Dans cet article, nous proposons de présenter les apports et limites des méthodes d'évaluation aléatoire appliquées aux expérimentations sociales. Nous nous appuyons sur un survol de la littérature internationale et sur notre propre expérience des méthodes d'expériences contrôlées appliquées à l'évaluation de programmes sociaux. La thèse sous-jacente se veut équilibrée. Il est clair que le développement d'évaluations aléatoires est un progrès indéniable, un élargissement de la boîte à outil des économistes qui s'intéressent aux questions sociales en France. Pour autant, ce n'est pas la fin de l'histoire de l'économie appliquée ni un nouveau *Gold Standard* méthodologique qui remplacerait toutes les autres approches, pour reprendre l'expression d'Angus Deaton (2009). Les méthodes expérimentales doivent occuper une place de choix dans l'ensemble des méthodes d'évaluation et il importe de bien connaître leur portée et leurs limites pour circonscrire cette place et les utiliser à bon escient. A cette fin, une première section décrit les grandes caractéristiques de ces méthodes, une deuxième section présente de façon équilibrée les principaux apports tandis qu'une troisième section est consacrée à une présentation nuancée de leurs limites.

L'évaluation aléatoire : de quoi parle-t-on ?

L'objectif de l'évaluation quantitative d'une expérimentation sociale est de mesurer les effets immédiats et différés d'un programme sur une ou plusieurs variables d'intérêt (par exemple, le taux d'accès à l'emploi, les sorties de la pauvreté monétaire, la fréquence des échecs scolaire, etc.). Il s'agit de vérifier si les effets attendus du programme ont bien été réalisés et aussi de savoir si le programme n'a pas eu d'autres effets, qui n'étaient pas nécessairement attendus. Pour mener à bien ce type d'étude, qui constitue le cœur de l'évaluation, il est

[5] Dans le cadre de l'expérimentation du RSA, ni la liste des départements expérimentateurs, ni le périmètre des zones tests dans chaque département, ni la liste des allocataires du RSA n'ont été choisi au hasard. Les départements, qui sont les véritables pilotes du RMI depuis la loi de décentralisation de décembre 2003, ont défini le périmètre des zones test sur la base d'une sélection raisonnée, en appliquant des critères et selon des contraintes qui leurs sont propres. A l'intérieur de ces zones, tous les allocataires du RMI bénéficient du rSa. L'expérimentation du RSA s'inscrit donc dans le registre des quasi-expériences. Chaque département réalise une expérience qui est au niveau national répétée plus de trente fois, selon des modalités qui varient à la marge. On est bien dans la situation la plus courante de l'évaluation des politiques publiques où les bénéficiaires de la politique ne font pas l'objet d'un tirage au sort. La définition des zones témoins a été adaptée en conséquence (Goujard et L'Horty, 2010).

indispensable d'adopter une approche quantitative, c'est-à-dire de mobiliser des données statistiques et des moyens de traitement adaptés aux caractéristiques des données. Cette étude d'impact va impliquer différentes étapes : le choix d'un protocole d'évaluation ; la construction d'un système d'information ; le traitement des données ; l'analyse des résultats, la rédaction d'une étude et de ses conclusions. L'objectif n'est pas seulement de construire des chiffres, il s'agit surtout de construire des chiffres de qualité, qui permettront de produire une mesure fiable et précise des effets du programme.

Nous pensons que la méthode de l'évaluation aléatoire appliquée à une expérimentation sociale peut se définir par la combinaison de quatre caractéristiques particulières. Si la présence d'un tirage au sort est la caractéristique la mieux connue, l'évaluation aléatoire suppose aussi la réunion de trois autres éléments qui chacun la distinguent de toute autre méthode d'évaluation : la présence d'un programme avec une dimension expérimentale qui va fournir l'objet de l'évaluation ; un partenariat particulier entre une institution expérimentatrice et une équipe d'évaluateurs ; la conception *a priori* d'un protocole d'évaluation et la construction *ad hoc* d'un système d'observation et de traitement de l'information.

Un tirage au sort

L'évaluation aléatoire d'une expérimentation sociale, dans sa version la plus élémentaire, implique de constituer deux groupes de personnes, puis à donner l'accès au dispositif que l'on souhaite évaluer à un groupe que l'on nommera le groupe test et à ne pas donner l'accès à l'autre groupe que l'on nommera le groupe témoin. Le point crucial est d'affecter au hasard les personnes éligibles dans chacun des deux groupes, en utilisant un tirage aléatoire simple dans une liste pré-constituée d'individus potentiellement éligibles au programme. Le tirage au sort peut porter sur des individus, sur des groupes, ou sur le traitement lui-même. Mais il doit nécessairement être mis en œuvre pour que l'évaluation fonctionne.

Pourquoi effectuer un tirage au sort ? Pourquoi ne pas se contenter de produire des statistiques descriptives sur la population qui bénéficie du programme et de les commenter ? La réponse est qu'un suivi d'indicateur dans le temps n'est pas une évaluation. Par exemple, il n'est pas satisfaisant de suivre l'évolution du taux d'accès à l'emploi des bénéficiaires d'un programme d'accompagnement, même en le comparant à celui des demandeurs d'emploi qui ne bénéficient pas du programme. De même, il n'est pas suffisant de suivre dans le temps la part de ménages pauvres parmi les bénéficiaires d'un dispositif expérimental de soutien aux bas revenus, même si on la compare aux ménages qui ne bénéficient pas du programme. La raison

est simple : les personnes qui ont bénéficié du programme social n'ont pas forcément les mêmes caractéristiques que celles qui n'en ont pas bénéficié. Les bénéficiaires du programme sont généralement sélectionnés sur la base d'un ensemble de caractéristiques qui ont des effets sur les trajectoires personnelles. L'échantillon des personnes qui vont bénéficier du programme social n'est donc pas représentatif de l'ensemble des éligibles. Pour qualifier ce phénomène, on parle de « biais de sélection ».

Les caractéristiques qui influencent les trajectoires et qui peuvent contribuer à la présence d'un biais de sélection sont de nature variées. Certaines sont observables. Par exemple, il existe de nombreuses études qui montrent que les chances de trouver un emploi lorsque l'on est au chômage baissent avec l'âge, augmentent avec la qualification, sont plus faibles pour les femmes, etc. Si le groupe de bénéficiaires comprend plus d'hommes, de jeunes, et de qualifiés, il va présenter un taux de retour à l'emploi mécaniquement plus élevé que celui des non bénéficiaires. Dans ce cas, le risque est alors de surévaluer les effets du programme. Si celui-ci est ensuite généralisé à l'ensemble de la population éligible, il ne produira pas les effets attendus. Ce type de biais sur les caractéristiques peut être convenablement corrigé avec des méthodes économétriques adaptées. Mais d'autres caractéristiques ne sont pas observables. Par exemple, la motivation, la carrière professionnelle antérieure, les accidents de vie, les loisirs et les pratiques culturelles, etc. peuvent exercer un effet sur les chances d'insertion mais constituent autant de variables qui ne peuvent pas être observées par l'évaluateur. Ces variables peuvent contribuer à un biais de sélection sans que l'on sache *a priori* dans quel sens joue ce biais.

Pour produire une mesure fiable des effets d'un programme, il est donc indispensable de s'affranchir de ces différents biais. La difficulté est de contrôler de l'hétérogénéité à la fois sur les variables observables et sur les inobservables. Pour y parvenir, il est nécessaire de mobiliser des techniques très sophistiquées qui requièrent un grand nombre d'observations, ou de construire des protocoles permettant de contrôler les données. Un tirage au sort est la meilleure façon pour que les personnes des deux groupes aient en moyenne les mêmes caractéristiques observables ou non observables. S'ils ont une taille suffisante (plusieurs centaines d'individus dans chaque groupe), les deux groupes auront exactement la même composition par âge, sexe, qualification, et aussi selon d'autres caractéristiques que l'on ne peut pas observer, par exemple la motivation, la capacité à coopérer avec les institutions, les opinions politiques, syndicales ou religieuses, etc. qui peuvent avoir un impact sur les variables d'intérêt. L'intérêt majeur d'une évaluation aléatoire est de s'affranchir de ces bais

de sélection sur les individus bénéficiant du programme de façon à produire un chiffrage très robuste avec une grande économie de moyen statistique ou économétrique.

Le tirage au sort est la réponse à ce que James Heckman (1998) qualifie de « problème de l'évaluation ». Pour mesurer les effets d'un programme, il faut idéalement pouvoir observer un même individu dans deux états du monde, celui où il bénéficie du programme et celui où il n'en bénéficie pas. Mais comment connaître ses réalisations, dans un état du monde qui ne s'est pas réalisé ? La réponse est de disposer d'un groupe de contrôle, appelé également contrefactuel. Dans l'évaluation aléatoire, ce contrefactuel a un statut particulier. Il est construit par le chercheur. Il n'est pas donné par la nature comme c'est le cas dans les quasi-expériences que l'on appelle aussi expérience naturelle. Dans les expériences contrôlées, le contrefactuel est construit et non fortuit, il est observé et non inventé.

Une expérimentation

Deuxième particularité majeure de l'évaluation aléatoire, il s'agit de mesurer les effets d'un projet innovant à dimension sociale, ou expérimentation sociale. Un dispositif expérimental est par définition limité à la fois dans le temps, il est temporaire, et dans l'espace, il est local. En France, la définition retenue pour l'expérimentation sociale[6] suppose que celle-ci doit être à la fois évaluable et généralisable. C'est d'ailleurs l'évaluation qui permet de savoir si l'expérience gagnerait à être généralisée, c'est-à-dire étendue dans l'espace, et prolongée, c'est-à-dire étendue dans le temps.

Le caractère innovant du programme renforce l'intérêt d'une évaluation par expérience contrôlée. En effet, l'évaluateur d'un dispositif innovant ne peut s'appuyer sur aucun précédent pour tenter d'inférer *a priori* les effets du programme. Comme le programme est souvent très original, il n'est pas toujours possible de mobiliser un cadre théorique formalisé pré-constitué pour analyser ses effets. On ne peut pas non plus mobiliser les enseignements d'une expérience naturelle. Certes, de nombreux programmes sociaux fournissent spontanément, de façon naturelle, un cadre qui se rapproche de celui d'une expérience

[6] Le site du Ministère de la Jeunesse et des Solidarités Actives précise que « l'expérimentation est une innovation de politique sociale initiée dans un premier temps à une échelle limitée, compte tenu des incertitudes existantes sur ses effets, et mise en œuvre dans des conditions qui permettent d'en évaluer les effets dans l'optique d'une généralisation ». (http://www.experimentationsociale.fr/)

contrôlée. Mais le stock d'expériences naturelles est néanmoins limité et l'on peut souhaiter disposer d'évaluations pour des mesures pour lesquelles aucune quasi-expérience n'est disponible. Ce sera le cas notamment de toutes les mesures radicalement nouvelles.

Un partenariat durable entre un évaluateur et un expérimentateur

Une troisième originalité de la méthode d'évaluation aléatoire, qui la distingue de toutes les autres méthodes, est qu'elle suppose la mise en œuvre d'un partenariat durable entre au moins trois acteurs : un expérimentateur, un évaluateur et un financeur. Cela fait d'une évaluation aléatoire une œuvre à la fois collective et inscrite dans la longue durée. Le partenariat débute en amont de la mise en œuvre de l'expérimentation et perdure jusqu'à la fin de l'évaluation.

Le financeur est l'Etat, un organisme public ou une fondation. Il s'agit d'une entité intéressée par la production de connaissances nouvelles et de portée générale sur les effets d'une innovation sociale. Puisque les résultats d'une évaluation s'apparentent à un bien public, il est naturel que son financement mobilise des budgets publics.

L'expérimentateur est une institution, par exemple une collectivité territoriale, un établissement public ou une association, qui envisage de mettre en œuvre une innovation à caractère social afin d'améliorer la situation des personnes dans son champ d'intervention. Elle est directement ou indirectement en contact avec des personnes qui vont bénéficier du dispositif innovant. L'expérimentation est localisée sur un terrain donné, mais pourrait être généralisée en cas de succès sur d'autres terrains et l'on s'interroge sur les conditions de cette généralisation. L'objectif peut être par exemple de lever un obstacle à l'accès à l'emploi, de mettre en place un nouvel instrument qui favorise l'insertion sociale ou économique, ou encore de combattre un déterminant supposé de l'exclusion sociale. Cette innovation peut correspondre à un très vaste ensemble de dispositifs : programme d'accompagnement au retour à l'emploi, dispositif de lutte contre l'échec en formation initiale ou continue, lutte contre des problèmes de santé, contre l'illettrisme, etc.… Dans ces domaines et dans tous les autres, il s'agit d'améliorer un dispositif existant ou de mettre en place un nouveau programme. L'innovation n'a pas encore été mise en œuvre et c'est en amont de l'expérimentation que la question de son évaluation est posée.

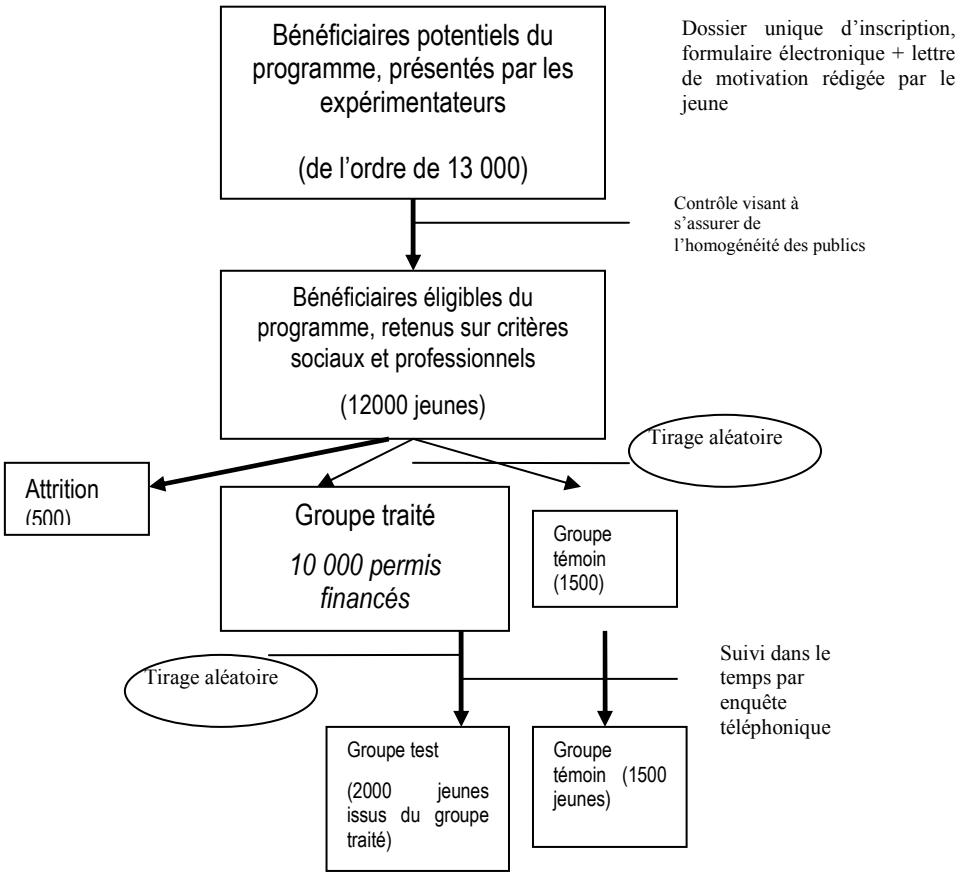
L'évaluateur est un centre de recherche ou un cabinet de conseil qui présente des références dans le champ des politiques sociales et de leur évaluation. Il maîtrise les technologies de l'évaluation adaptées à la problématique de l'expérimentation. L'évaluation est effectuée de

façon externe et indépendante de l'expérimentateur. Cela n'exclut pas, bien au contraire, que l'expérimentateur mobilise des outils internes d'observation sur le déroulement et les effets de son action. Mais ces outils ne se substituent pas à l'évaluation externe et indépendante du programme. L'évaluateur n'évalue pas l'expérimentateur. Il évalue l'expérimentation. Par ailleurs, les résultats de l'évaluation ne sont en aucun cas donnés par avance. Cela fait de l'expérimentation et de son évaluation une activité risquée, qui implique des coûts certains pour des gains hypothétiques. En particulier, les conclusions de l'évaluation peuvent amener à renoncer à la généralisation de l'expérimentation.

Un protocole d'évaluation et un système d'observation et de traitement des données

Enfin, une évaluation aléatoire requiert qu'un protocole soit construit en amont de l'expérimentation. Ce protocole prend la forme d'un schéma qui décrit le déroulement de l'expérimentation et de son évaluation. Il précise les modalités de constitution des groupes test et témoins, les effectifs de chaque groupe en tenant compte de l'attrition éventuelle durant l'expérimentation, le déroulement des interrogations ou des enquêtes qui vont permettre de collecter des données tout au long de l'expérimentation. A chaque évaluation correspond un protocole spécifique en fonction des particularités de l'expérimentation. A titre illustratif, le schéma 1 donne le protocole retenu dans le cadre du projet « 10 000 permis pour réussir » qui consiste à évaluer les effets d'un financement et d'un accompagnement de jeunes en difficulté d'insertion pour les aider à passer leur permis de conduire. L'évaluation vise à mesurer les effets de la subvention et des dispositifs d'accompagnement sur un certain nombre de variables d'intérêt qui indiquent la prise d'autonomie des jeunes (insertion professionnelle, autonomie résidentielle, mise en couple, etc.). La première étape consiste à établir une liste d'éligibles au programme sur la base de critères d'âge et de situation sociale. Les jeunes éligibles sont présentés par les expérimentateurs et remplissent un dossier unique d'inscription qui prend la forme d'une application extranet. Un premier tirage au sort a lieu pour déterminer si le jeune va effectivement recevoir la subvention et l'accompagnement (groupe traité) ou s'il n'aura pas accès à cette aide mais uniquement au droit commun (groupe témoin). Le protocole est construit pour que le groupe traité comprenne *in fine* 10 000 jeunes. Au sein de ce groupe traité, un échantillon est tiré et fait l'objet d'enquêtes téléphoniques six mois puis douze mois après le début de l'expérimentation. Les informations portent sur les variables d'intérêt et sont ensuite comparées à celles du groupe témoin recueillies de la même manière.

Schéma 1. Un exemple de protocole : le projet « 10 000 permis pour réussir »



Dans la construction du protocole, il est nécessaire de distinguer les mesures auxquelles les individus du groupe traités ont accès. En effet, si ces individus bénéficient simultanément de plusieurs mesures, il sera impossible d'évaluer et de comparer leur efficacité respective. Pour tester l'efficacité de plusieurs mesures prises individuellement ou de façon combinée, il existe des protocoles d'expérience contrôlée plus sophistiqués, avec trois ou quatre groupes test, avec des entrées différées dans le dispositif ou avec plusieurs sélections successives dans les groupes [7]. Parfois, c'est le programme lui-même qui fait l'objet d'un tirage au sort, avec par exemple des traitements différenciés selon les personnes [8] . La seule constante est qu'il est nécessaire d'avoir au moins une étape de l'expérience où un tirage au sort a lieu. Seul le

[7] Par exemple si on souhaite évaluer les effets propres et cumulés de deux mesures, on peut constituer quatre groupes : un groupe qui bénéficie seulement de la mesure 1, un groupe qui bénéficie seulement de la mesure 2, un groupe qui bénéficie des deux mesures, un groupe qui ne bénéficie d'aucune des deux mesures

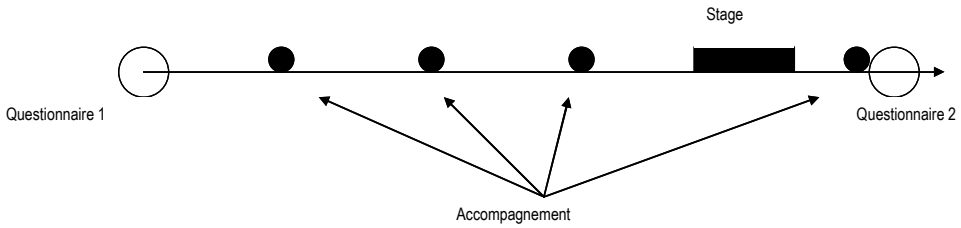
[8] Par exemple, les individus du groupe test bénéficient d'une subvention dont le montant est déterminé par tirage au sort.

hasard permet de distinguer les individus selon un processus parfaitement indépendant de leurs caractéristiques personnelles, qu'elles soient observables ou non observables.

La définition du protocole suppose aussi de préciser les modalités de recueil des observations durant toute l'expérimentation. On combine ici l'apport de fichiers de gestion, tels qu'ils sont utilisés par l'institution expérimentatrice, et celui d'enquêtes statistiques spécifiques qui permettent de compléter l'information des fichiers de gestion sur les variables de contrôle et les variables d'intérêt. C'est un véritable plan d'enquête et d'appariement des sources qu'il faut donc concevoir[9]. En pratique, une étape préalable de l'évaluation va donc consister à expertiser le système d'information qui va permettre de mesurer les variables de contrôle et de suivre les variables d'intérêt dans le temps. Il est nécessaire de réaliser un travail spécifique d'évaluation du système d'information de l'expérimentateur et des moyens de le compléter. Il faut se donner des outils d'observation, construire des indicateurs, recueillir des données et déployer des technologies de traitement pour ces données. Par exemple, on aura besoin de savoir pour chaque individu sa date d'entrée dans le programme, la durée de séjour dans le dispositif, et des éléments sur sa situation après la sortie du programme. On aura besoin de compléter les informations contenues dans le fichier de gestion par des enquêtes statistiques réalisées auprès de bénéficiaires et aussi auprès de non bénéficiaires de l'expérimentation. Eventuellement, il sera nécessaire de répéter ces enquêtes complémentaires plusieurs fois afin de suivre les populations dans le temps. Par exemple, dans le cadre de l'évaluation du programme « Tous en stage », qui consiste à accompagner des élèves de 3 ème qui résident dans des zones urbaines sensibles pour les préparer au stage obligatoire de découverte de l'entreprise, on réalise une enquête par questionnaire avant le début de l'accompagnement puis une deuxième enquête après le stage (schéma n°2). Les enquêtes ont lieu à la fois dans des classes qui bénéficient de quatre séances d'accompagnement et dans celles qui n'en bénéficient pas, ce qui permet de réaliser une évaluation par double différences (avant/ après dans les deux types de classes).

[9] La construction de ce système d'observation *ad hoc* requiert une mise en conformité avec la règlementation en vigueur dans le domaine de l'informatique et des libertés ce qui implique de réaliser des déclarations auprès de l'autorité compétente, la CNIL.

Schéma 2. Un exemple de plan d'enquête : le projet « Tous en Stage »

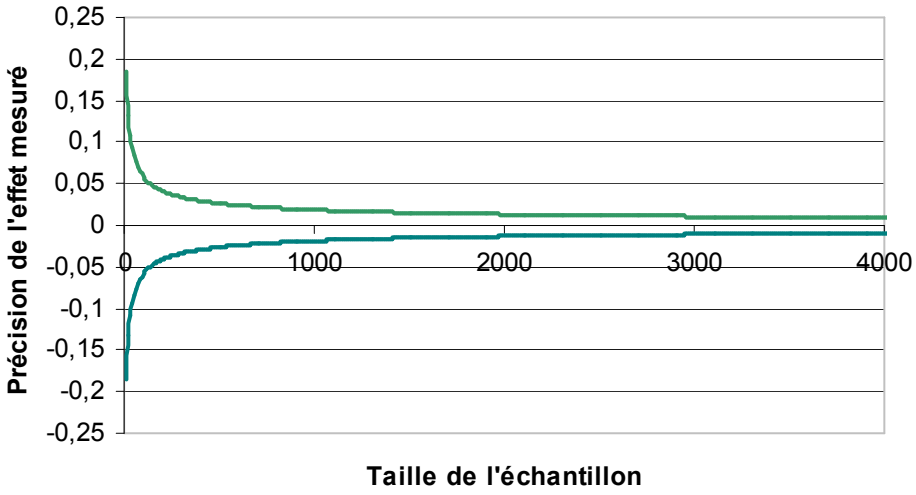


Le chercheur doit aussi définir la taille optimale des échantillons test et témoin, qui résulte d'un arbitrage entre la précision souhaitée et le coût de production de l'information. La précision d'une enquête, ou puissance statistique, augmente de façon décroissante avec le nombre d'observations. Le coût de l'enquête augmente lui aussi de façon décroissante avec le nombre d'observations, mais dans une moindre mesure. Il existe ainsi une taille optimale de l'enquête au-delà de laquelle le gain marginal de précision est plus que compensé par le coût marginal de collecte. Le chercheur doit également faire le choix d'une méthode de traitement des observations. Pour un nombre d'observations données, c'est-à-dire à coût de l'enquête donné, il est possible d'accroître la puissance avec des techniques spécifiques. L'intuition est d'améliorer la qualité du contrefactuel, c'est-à-dire de rendre le groupe témoin encore plus comparable au groupe test. Pour Bruhn et Mac Kenzie (2009), il s'agit d'arbitrer entre trois grandes familles de méthode qui sont la stratification, l'appariement et la re-randomisation. Ces différentes méthodes produisent des résultats analogues pour des échantillons suffisamment larges, de plus de 300 personnes. Mais pour capter des effets faibles, sur des variables très inertes ou lorsque l'on dispose de moins d'observations, la stratification et l'appariement sont préférables.

Dans l'illustration précédente tirée du projet 10 000 permis pour réussir, nous proposons le chiffre de 2000 jeunes pour le groupe test et le groupe témoin car nous souhaitons mesurer avec précision les effets de la subvention et de l'obtention du permis de conduire sur le taux de retour à l'emploi. D'une part, le taux d'échec au permis de conduire des jeunes peut être élevé (il est d'un peu plus de 40 % en 2004 pour l'ensemble de la population). D'autre part, les études américaines indiquent un ordre de grandeur pour l'effet de la possession d'une voiture sur les chances d'être en emploi de 15 % (Gurley T. et Bruce D. [2005], Ong P. [2002], Raphael S. et Rice L. [2002]). Si l'on prend comme référence une probabilité d'accès à l'emploi de 10 % tous les six mois, il est important d'avoir un échantillon assez large pour

déceler un choc de 15 % sur cette probabilité, soit une hausse de 1, 5 point. Avec un échantillon de 2000 individus, nous pourrons déceler un choc de 1,3 point sur la probabilité d'accès à l'emploi, pour une probabilité de départ de 10 % (cf graphique ci-dessous). Il est difficile d'être plus précis en l'absence d'une information détaillée sur le nombre des expérimentateurs, les effectifs traités par chaque expérimentateur, la probabilité de succès au permis de conduire pour les jeunes en difficulté d'insertion, et les chances d'accès à l'emploi de ces catégories particulières de jeunes, avec et sans permis de conduire.

Graphique 1. Précision de l'effet mesuré en fonction de la taille de l'échantillon



Assignation aléatoire des personnes dans un groupe test et un groupe témoin, démarche expérimentale, relation partenariale entre un expérimentateur et un évaluateur externe et indépendant, mise au point *a priori* d'un protocole d'évaluation et d'un système d'observation et de traitement des données, constituent les principales spécificités de l'évaluation aléatoire d'une expérimentation sociale. Elles sont suffisamment singulières pour accorder à l'évaluation aléatoire une place spécifique dans l'ensemble des méthodes d'évaluation.

Les apports de l'évaluation aléatoire

La présentation que nous venons de retenir, qui définit l'évaluation aléatoire à partir de quatre grandes caractéristiques, peut être conservée si l'on souhaite décrire les principaux apports de cette approche. C'est l'objet de cette section où nous montrons que l'assignation aléatoire apporte un chiffrage d'une qualité sans doute inégalée par les autres techniques d'évaluation, que la dimension expérimentale permet un ciblage très fin de l'objet de l'évaluation, que le partenariat est créatif et que l'existence d'un protocole et la conception d'un système d'observation élargit le domaine de maîtrise de l'évaluateur.

Assignation aléatoire : aucune autre alternative n'est pleinement satisfaisante

La conférence pour le prix Nobel que James Heckman a donné en 2000 a été entièrement consacrée à la problématique de l'évaluation des politiques publiques. Il y distingue deux grandes familles de méthode d'évaluation quantitative qui se différencient par la nature des données utilisées. La première famille est celle des évaluations qui utilisent des données d'expériences contrôlées (les évaluations aléatoires). La deuxième famille correspond aux évaluations qui mobilisent des données non expérimentales mais qui tentent de se rapprocher des conditions des données expérimentales (on parle aussi de données de quasi-expérience, ou d'expérience naturelle).

Ces deux grandes familles se hiérarchisent du point de vue de la qualité du chiffre qu'elles parviennent à produire. Les évaluations les plus précises et les plus fiables sont obtenues avec des données d'expériences contrôlées et aléatoires. Aux yeux des spécialistes les plus exigeants, ces méthodes sont les seules qui permettent de véritablement prouver les effets d'une expérimentation. La qualité de la preuve est en quelque sorte équivalente à celle d'un flagrant-délit pour une cour de justice, alors que les autres méthodes, qui mobilisent des données non expérimentales, reviennent à accumuler des présomptions de culpabilité sans jamais véritablement administrer la preuve de façon définitive. Certes, elles réduisent les marges d'erreur, parfois de façon très significative, mais elles sont toujours moins fiables qu'une évaluation aléatoire. C'est la raison pour laquelle la bonne démarche évaluative est de chercher en premier lieu à construire une expérience contrôlée et à se replier sur une méthode non contrôlée en cas d'impossibilité.

S'il est impossible de conduire une expérimentation sociale qui intègre une évaluation selon un protocole randomisé, on peut avoir recourt à une évaluation sur des données non

expérimentales. Une simple comparaison des moyennes des variables de performance des individus bénéficiant de la mesure et de ceux qui n'est bénéficient pas serait toutefois incorrecte, du fait de l'existence des biais de sélection. Il convient dans tous les cas d'appliquer des techniques statistiques et économétriques permettant de corriger au mieux ces biais, et pour ce faire de disposer d'échantillons de taille importante[10].

Idéalement, il conviendrait de comparer la variable de performance d'un même individu selon qu'il bénéficie ou non d'une mesure ou encore d'un programme, dont l'efficacité est à tester. Or, par définition, soit un individu donné en bénéficie, soit il n'en bénéficie pas. Dans l'un ou l'autre cas, on observe donc une seule valeur de la variable de performance pour chaque individu. Il faut donc estimer la variable de performance d'un individu bénéficiant de la mesure s'il n'en bénéficiait pas, et la variable de performance d'un individu ne bénéficiant pas de la mesure s'il en bénéficiait.

Pour y parvenir, la méthode la plus utilisée aujourd'hui est la « méthode d'appariement » qui a été proposée initialement dans le champ de l'économie de l'éducation par D. Rubin (1974). L'intuition est la suivante. Si deux individus ont la même probabilité de bénéficier de la mesure, et que le premier l'est alors que le second ne l'est pas, tout se passe comme si l'accès à la mesure était affecté aléatoirement entre ces deux individus. Dans ce cas, la comparaison des variables de performance de ces deux individus fournirait une estimation de l'effet de la mesure sur la variable de performance. Pour estimer l'effet de la mesure, on considère les individus qui en bénéficient et, pour chacun d'entre eux, on estime ce que serait leur variable de performance en l'absence de la mesure. Pour cela, on sélectionne un individu bénéficiant de la mesure et on effectue la moyenne des variables de performance des individus n'en bénéficiant pas qui ont la même probabilité que lui d'en bénéficier. On effectue ensuite la moyenne des écarts obtenus sur tous les individus bénéficiant de la mesure. On effectue une opération symétrique pour estimer l'effet potentiel qu'aurait la mesure sur les individus n'en bénéficiant pas, et on estime l'effet total en appariant chaque individu avec ceux du groupe auquel il n'appartient pas.

Le principe de cette méthode consiste donc à comparer la variable de performance de chaque individu bénéficiant de la mesure, avec celle d'un individu « clone » ne bénéficiant pas de la

---

[10] Pour une présentation complète de ces techniques, voir Angrist J. et A. Krueger, (1999), Blundell R. et Costa Dias M. (2000), Dehejia R. et Wahba S., (1999) ou Meyer B. (1995).

mesure. On considère que ces deux individus sont « clones », autrement dit qu'ils ont les mêmes caractéristiques dans la mesure où ils ont la même probabilité de bénéficier de la mesure. Apparaît ici une limite importante de cette méthode. Si un petit nombre de caractéristiques influencent la probabilité de bénéficier de la mesure, on considérera comme « clones » deux individus qui dans les faits auront un grand nombre de caractéristiques distinctes, et de fait, on attribuera à la mesure des effets qui seront en partie liés à ces différences de caractéristiques non contrôlées.

On le voit, les méthodes alternatives à l'évaluation aléatoire sont très sophistiquées sur le plan des techniques statistiques. Elles requièrent d'être mises en œuvre par des experts confirmés et ne garantissent pas l'absence de tout biais. Surtout, elles supposent pour être mises en œuvre de pouvoir disposer d'un grand nombre d'observations, ce qui n'est pas toujours compatible avec une expérimentation sociale qui porte le plus souvent sur des échantillons de quelques dizaines ou quelques centaines de bénéficiaires. Pour mener à bien une évaluation quantitative permettant de produire une mesure fiable des effets d'un programme social, il n'y a donc guère d'alternative au choix d'un protocole d'évaluation aléatoire lorsque l'expérimentation est de petite taille.

En produisant un contrôle *a priori* des biais de sélection, les méthodes d'évaluation aléatoire permettent de faire l'économie des multiples traitements de l'économétrie de l'évaluation dont l'objectif est de tenter de corriger des biais de sélection sur les observables ou sur les inobservables. L'économétrie peut alors se recentrer sur des aspects plus intéressants pour l'économiste. Par exemple, plutôt que de proposer une évaluation au point moyen, on peut analyser toute la distribution des effets du traitement. On peut tenter de mesurer les effets conditionnels de différentes variables sur toute la distribution des variables d'intérêt.

Expérimentation : un ciblage potentiel très précis de l'évaluation

L'avantage lié à la dimension expérimentale du programme réside dans la possibilité donnée au chercheur de cibler l'objet de son évaluation de façon à la fois souple et extrêmement précise. Il devient techniquement possible d'évaluer n'importe quelle composante d'un programme social. On peut estimer les effets d'un changement dans les conditions d'éligibilité du programme (l'ouverture à des nouveaux publics ou la restriction d'accès de certains publics), dans la durée de séjour dans le programme, dans les modalités d'accompagnement des personnes, dans les différents leviers qui agissent sur l'incitation des

personnes à participer au programme, notamment les leviers monétaires ou non monétaires, dans l'information donnée aux participants, etc. On n'évalue donc pas nécessairement un programme dans son ensemble mais parfois une réforme dans un programme existant, tel un changement dans la valeur de l'un des paramètres d'un dispositif social innovant. L'objet est toujours une innovation de *process* dans un dispositif social expérimental, mais cette innovation n'est pas nécessairement radicale, il peut s'agir d'une innovation incrémentale au sein d'un programme. On peut d'ailleurs comparer les effets d'un petit choc sur l'un de ces paramètres à l'effet d'un choc de plus grande ampleur de façon à repérer des non linéarités éventuelles dans les comportements des bénéficiaires. Les possibilités d'évaluation apparaissent donc très vastes et ne sont bridées que par l'imagination de l'évaluateur et de celle de l'expérimentateur.

Le champ des possibles s'élargit encore lorsque l'on considère la possibilité d'évaluer simultanément plusieurs composantes d'un programme. Toutes les mesures qui viennent d'être données à titre d'illustration, peuvent être évaluées de façon isolées ou de façon combinées, afin d'identifier des effets joints. L'idée peut être de vérifier si des mesures sont simplement additives ou si elles produisent des synergies et entretiennent des relations de complémentarité qui renforcent leurs effets. On peut aussi vérifier si les mesures ne sont pas substituables les unes aux autres. On évalue en quelque sorte des multi-thérapies.

Cette extrême flexibilité des méthodes aléatoires n'est pas une caractéristique partagée par les méthodes d'évaluation non expérimentales. Dans les quasi-expériences, le chercheur ne peut pas manipuler son objet d'étude, qu'il prend comme une donnée. Dans les méthodes d'évaluation *ex ante*, à l'aide de simulation de modèles théoriques pré-construits, on n'a pas la possibilité en générale d'évaluer la linéarité des effets de différentes mesures.

Partenariat : une co-production créatrice

Pour Esther Duflo, le principal intérêt des méthodes aléatoires ne réside pas tant dans la correction des biais de sélection ou dans la flexibilité donnée par la démarche expérimentale. Il est lié à la particularité de la relation qui se noue dans la durée entre l'expérimentateur et l'évaluateur. Ce partenariat original est un processus dynamique et itératif qui fait avancer à la fois la recherche et l'action publique. Il est créatif dans la mesure où il peut produire de façon intrinsèque des connaissances nouvelles.

L'expérimentateur participe à l'évaluation en mettant à disposition son propre système d'observations (données de gestion). L'évaluateur devient quant à lui un co-expérimentateur, il va être en position de contribuer au *design* du programme social. Il n'est donc plus dans un rôle passif vis-à-vis de la politique publique, alors qu'il la prend comme une donnée dans les autres formes d'évaluation. Si cette position nouvelle peut effectivement favoriser une créativité particulière, il est clair que l'évaluation doit demeurer externe et indépendante pour rester crédible.

Protocole et système d'observation : une nouvelle relation avec la théorie et avec les données

L'évaluation aléatoire ne modifie pas uniquement la position du chercheur par rapport aux politiques qu'il souhaite évaluer, elle altère également sa position vis-à-vis de la théorie et vis-à-vis des données. Dans chacun de ces domaines, la responsabilité du chercheur est étendue.

Vis-à-vis de la théorie, les méthodes d'évaluation aléatoire peuvent rendre de précieux services en testant la validité de certaines hypothèses, en estimant des paramètres structurels très fins pour lesquels on ne disposent d'aucune évaluation (par exemple, l'élasticité prix de la demande de soins pour les faibles revenus), ou encore, en testant les prédictions issues de modèles. Bref, sur de multiples plans, la conception des protocoles peut être mise au service du perfectionnement des théories. « *The goal is better theory* » écrivent Banerjee et Duflo [2008], sans pour autant que l'on puisse réduire l'apport potentiel de ces méthodes à ce soutien à la théorie. On peut par exemple tester des schémas de conditionnalités sur des traitements qui seraient *a priori* inaccessibles à la théorie seule. On peut aussi produire des résultats sur les effets de certains traitements qui ne sont pas dérivables d'un modèle théorique. L'exemple le plus cité en économie du développement est le travail de Miguel et Kremer (2002) sur l'absentéisme des élèves des écoles du Kenya dans lequel les traitements contre les vers intestinaux produisent les effets les plus importants lorsqu'ils sont massifs, du fait d'externalités locales.

Vis-à-vis des données, la position du chercheur est là aussi étendue. Il doit expertiser le système d'observation existant et le compléter s'il y a lieu par un dispositif *ad hoc* de recueil d'information. Ce faisant, le chercheur à désormais un rôle actif dans la production des données, qui ne lui sont plus « données » mais qui sont construites pour les besoins de l'évaluation, ce qui élargit son domaine de responsabilité. Cela ouvre la possibilité de

contourner les défauts habituels des appareillages d'observation quantitatifs qui ne contiennent pas assez d'observations préalables à la mise-en œuvre de la politique, qui ne contiennent pas toujours les bonnes variables et ne portent pas nécessairement sur les publics les plus intéressants (un fichier de gestion porte par définition sur les publics gérés par une institution, et ne couvre ni sur les sortants du dispositif, ni ceux qui ne sont pas rentrés dans le dispositif). Le fait d' « avoir la main » sur la construction des données permet en outre de produire des observations pour les *outcomes* jugés pertinents pour l'objet de recherche et de commencer à les recueillir avant que la politique ne soit implémentée.

Au total, on voit que l'évaluateur qui déploie une méthode aléatoire devient à la fois producteur de données et co-producteur d'un programme. Son métier change. Il est beaucoup plus sur le terrain et beaucoup moins « derrière son ordinateur à faire tourner ses programmes ». En partie du fait de cette implication accrue du chercheur, l'évaluation expérimentale est beaucoup plus transparente pour l'expérimentateur et pour le financeur. Elle est alors très lisible pour le *policy maker* qui va souvent lui accorder davantage de crédibilité que d'autres approches qui ne partagent pas ces propriétés.

Les limites de l'évaluation aléatoire

Toute médaille a son revers. Dans cette section, nous conservons le même ordre de présentation pour décrire les principales limites des évaluations aléatoires en les reliant à chacune des quatre grandes caractéristiques des évaluations aléatoires.

La question éthique du tirage au sort

Même si le tirage au sort est le meilleur moyen de garantir *ex ante* l'égalité des chances, le fait de séparer la population éligible en un groupe test et un groupe témoin se traduit par une inégalité *ex post* de situation. Certes, cette rupture au principe d'égalité se justifie par le caractère temporaire de l'expérimentation et par la perspective de sa généralisation ultérieure en cas de succès. Mais elle n'en reste pas moins contraire au principe d'égalité de traitement des personnes, ce qui pose une question éthique. De surcroît, un protocole d'expérimentation aléatoire implique de priver une partie de la population des ressources qui pourraient lui être nécessaires pour améliorer sa situation. Si ces ressources ont effectivement un effet sur la trajectoire des personnes, leur privation pour le groupe témoin peut aller à l'encontre de

l'objet même de l'institution expérimentatrice dont la finalité première est d'améliorer le bien-être de personnes en difficulté d'insertion sociale et/ou économique.

La question éthique est soulevée par les travailleurs sociaux qui sont en contact direct avec les publics traités ou encore par les élus qui souhaitent naturellement que le plus grand nombre de bénéficiaires aient un accès immédiat à l'innovation sociale. Il est souhaitable qu'elle soit prise en considération dans toutes les expérimentations sociales, au cas par cas, quel que soit leur protocole d'évaluation, et plus encore lorsqu'est envisagée une évaluation aléatoire. Parfois même, cette réflexion pourra conduire à abandonner l'idée d'une évaluation par tirage au sort. C'est le cas lorsque les expérimentations mettent en jeu des ressources vitales ou déterminantes pour la trajectoire de vie des personnes. Le problème se pose notamment dans de nombreux programmes d'aides mis en œuvre dans les pays en développement. S'agissant des expérimentations sociales en France, on se dit qu'il n'est pas envisageable d'organiser à des fins expérimentales une privation de ressource pour des personnes en difficulté d'insertion sociale. Pour cette raison, il s'agit d'appliquer les méthodes d'évaluations aléatoires uniquement sur des expérimentations où l'on veut tester un « plus » par rapport au droit commun. Le groupe témoin dispose de l'accès aux ressources de droit commun. Le groupe test dispose de surcroît d'un accès à un complément de ressource, à un service amélioré, dont il s'agit de tester l'efficacité. Lorsqu'il s'agit de tester les effets d'un surcroît de service, d'un supplément au droit commun, l'expérimentation ne soulève plus les mêmes obstacles éthiques.

En outre, le tirage au sort peut être considéré comme un moyen plus juste, plus transparent et plus légitime que l'affectation arbitraire d'une ressource rare au sein d'une population éligible. Dans tous les cas, l'accès au programme est généralement limité par les moyens humains ou financiers consacrés à l'expérimentation, le tirage au sort est alors une solution pour donner un accès sélectif au programme. Elle est d'ailleurs parfaitement compatible avec l'affectation selon une sélection raisonnée, en mobilisant des critères sociaux visant à prioriser des publics cibles. Pour combiner les deux modes d'affectation, il suffit en effet de définir une liste de personnes éligibles au programme en mobilisant ces critères, puis d'effectuer un tirage au sort au sein de cette liste de façon à constituer le groupe test et le groupe témoin.

De surcroît, il est souvent possible en pratique de modifier les paramètres de l'expérimentation et de son évaluation, afin de lever les obstacles à la mise en œuvre d'un protocole randomisé. Par exemple, on peut mettre en place un dispositif de « joker » qui

permet à l'expérimentateur d'exclure du tirage au sort une personne éligible pour lui donner un accès automatique au traitement. On peut aussi envisager de réduire la taille du groupe témoin si la puissance statistique est suffisante (la probabilité d'appartenir au groupe test peut être supérieure à 1/2). On peut également reporter dans le temps l'accès au programme social. Le groupe test est alors constitué des personnes qui ont un accès immédiat au programme. Le groupe témoin est constitué de ceux qui auront un accès différé, par exemple six mois plus tard. Si des différences sont observées entre les deux groupes du point de vue des variables d'intérêt, elles pourront être attribuées au fait d'être effectivement passé par le programme. Ce type de protocole garantit que tous les éligibles auront effectivement accès au dispositif, ce qui est souvent une attente légitime de l'expérimentateur.

Par ailleurs, si le tirage au sort ne peut être effectué sur les personnes, il est également possible de randomiser le traitement. Par exemple, dans un protocole où l'on évalue l'accès à des dispositifs de micro-crédits dans des villes moyennes, où il n'est pas possible d'affecter au hasard la possibilité de bénéficier du programme, il est envisageable de tenter de moduler de façon aléatoire les conditions du crédit, son coût, sa durée, les conditions de garanties, etc. ce qui permettra *in fine* de réaliser une évaluation satisfaisante sur le plan de la rigueur scientifique.

Plutôt que de sélectionner au hasard des personnes dans des listes, une autre solution peut être de tirer des groupes de personnes au hasard. Dans de nombreuses expérimentations, on effectue ainsi un tirage au sort de classes d'école, d'établissements ou encore, de localités. La rupture au principe d'égalité est plus acceptable si elle est collective, surtout si le tirage des groupes est combiné à un traitement différé (les groupes témoins d'aujourd'hui sont les groupes traités de demain). Au-delà de cet aspect éthique, le tirage de groupes présente d'autres avantages. Dans les manuels de théorie des sondages, ce type de tirage est appelé « tirage en grappes » (Gouriéroux, 1981). C'est un cas particulier d'un plan de sondage à deux degrés dans lequel les unités primaires font l'objet d'un tirage aléatoire simple et dans lequel les unités secondaires (en l'occurrence, des personnes) sont toutes interrogées (le deuxième tirage est un recensement). Le tirage en grappe ne requiert pas l'établissement d'une liste de tous les éligibles qui soit exhaustive et préalable à l'expérimentation et évite le risque d'un

échantillon aberrant. Il peut même améliorer la puissance statistique si les unités primaires sont suffisamment hétérogènes (i.e. si la variance inter-groupe est forte)[11].

Ces exemples illustrent le fait que la prise en compte des questions éthiques n'est pas incompatible avec le montage d'évaluation aléatoire même si elle impose un effort de réflexion commune à l'expérimentateur et à l'évaluateur. Il s'agit bien souvent de trouver des solutions astucieuses à des questionnements et à des situations toujours particulières. Il n'en reste pas moins que tous les programmes sociaux ne sont pas évaluables selon une comparaison entre un groupe test et un groupe témoin avec assignation aléatoire des personnes dans les deux groupes.

La constitution d'un groupe test et d'un groupe témoin implique une autre limite bien connue : le risque de voire les personnes changer de comportements parce qu'elles sont traitées (on parle alors d'effet Hawthorne) ou parce qu'elles sont témoins (on parle d'effet John Henry). Mais ce type d'effets, bien réel, n'est pas lié à la dimension aléatoire de l'évaluation. Il est présent dès que l'on met en œuvre une politique catégorielle avec des personnes qui en bénéficient et d'autres qui n'en bénéficient pas, que la politique soit expérimentale ou non, et avec ou sans tirage au sort.

Expérimentation et effets d'équilibre

Les expérimentations sont circonscrites à la fois dans l'espace et dans le temps, mais l'on espère en tirer des enseignements en vue d'une généralisation. Ces deux caractéristiques ne sont pas toujours compatibles. Un changement d'échelle temporelle ou spatiale peut produire un changement dans les effets d'un programme social. Les mécanismes mettent en jeu ce que l'on appelle des effets d'équilibre. L'extension d'un micro-dispositif a des effets agrégés qui modifient les équilibres de marché et les prix. Au travers de ces effets, le traitement a un impact sur les non traités.

Ces effets d'équilibre jouent dans un sens a priori indéterminés. Dans certains cas, les effets agrégés sont plus importants que l'agrégation des micro effets (par exemple, dans le cas des

[11] La stratification est un cas symétrique où les unités primaires sont toutes interrogées (recensement) et où il y a un tirage au sort des unités secondaires. Elle peut améliorer la puissance statistique par rapport à un tirage à un seul degré si les unités secondaires sont suffisamment hétérogènes (i.e. si la variance inter-groupe est faible). Pour une présentation du problème et d'une solution, la macro-cube de l'INSEE, voir Lopez et Rouaud [2010].

traitements contre les vers intestinaux étudiés par Miguel et Kremer, 2002). Dans d'autres cas, que l'on pense être les plus fréquents sans pouvoir véritablement l'affirmer, les effets d'un programme généralisé sont atténués par rapport aux effets mesurés au seul niveau local. On peut même imaginer aussi que des effets d'équilibre positifs soient strictement compensés par des effets négatifs et que le tout soit exactement la somme des parties. On retiendra que la mesure des effets d'un traitement local n'est pas nécessairement la mesure des effets d'un traitement généralisé. Dés lors, "étendre les résultats d'une étude expérimentale aux réformes qui couvrent une économie toute entière, est tout simplement impossible (Rodrik, 2008).

L'argument des effets d'équilibre peut être retourné contre ceux qui l'emploient : une évaluation macro-agrégée sans dimension micro-économique est affectée par des effets de composition et des biais de sélection. Les résultats agrégés ne peuvent être présents sur aucun des terrains où la mesure est mise en œuvre. Surtout, on se dit que le seul moyen de mesurer ces effets d'équilibre est de développer des expérimentations à différentes échelles, en particulier des expérimentations de taille intermédiaire ou à différentes niveaux. La solution n'est pas moins d'évaluation aléatoire mais davantage (Banerjee et Duflo, 2008).

Partenariat et biais de terrain

L'évaluation aléatoire d'une expérimentation sociale peut aussi être confrontée à l'existence de biais de sélection dans le choix des terrains d'expérimentations et des institutions expérimentatrices. Lorsqu'une expérimentation est conduite sur un territoire spécifique par un expérimentateur spécifique, il est possible qu'elle produise des effets spécifiques, qui ne sont pas généralisables. Ainsi, à l'origine d'une expérimentation se trouve une équipe d'expérimentateurs potentiels porteurs d'un projet. Par nature, ce type d'institution est favorable à ce type de protocole et est prêt à fournir les efforts nécessaires à sa mise en œuvre. Leur motivation à la réussite de l'expérimentation et à la démonstration d'effets significatifs de celle-ci peut également conduire les expérimentateurs à introduire plus ou moins consciemment des mesures d'accompagnement dans le groupe test qui ne sont pas prévues dans le protocole. Les effets de la mesure testée sur le groupe test peuvent s'en trouver influencés. Le risque est alors que les évaluations aléatoires accumulent des faits et non de la connaissance, pour reprendre le reproche formulé par Angus Deaton (2009). Ce risque était déjà souligné dans l'article de synthèse de Heckman (1992).

La dépendance au contexte local est un fait avéré dont il faut tenir compte. Ce sujet n'est d'ailleurs pas pris en compte dans les études purement macro, comme le soulignent Banerjee et Duflo (2008). Néanmoins, l'existence de ce biais de terrain, comme c'était le cas aussi pour les effets d'équilibre, ne plaide pas pour réaliser moins d'expérimentation, mais pour en réaliser davantage. Une façon de s'affranchir de ce biais de sélection consiste en effet à conduire simultanément l'expérimentation sur des sites différents, avec des expérimentateurs différents. La sélection des sites s'effectuant si possible par tirage au sort. Les mêmes mesures seront alors testées en mobilisant les mêmes instruments et en suivant les mêmes protocoles, de sorte que toute différence d'effet de ces mesures entre les sites sera imputable aux spécificités de ces derniers. Tel est par exemple le cas du projet 10 000 permis pour réussir qui comprend une centaine d'expérimentateurs de nature différente (Missions Locale, conseils généraux, auto-écoles sociales, etc.) répartis sur l'ensemble du territoire,

Protocole et système d'observation : le coût organisationnel et logistique de l'évaluation aléatoire

Une dernière catégorie d'obstacle à la mise en œuvre d'expérimentations et à leur évaluation selon des protocoles exigeants peut être évoquée. Il s'agit des difficultés matérielles rencontrées par les opérateurs de ces programmes. Parce qu'une expérimentation est toujours innovante, elle bouscule les routines organisationnelles, les façons de faire, les pratiques habituelles des institutions. Dès lors, la réalisation d'une innovation sociale confronte bien souvent l'expérimentateur à de multiples sources de rigidités insoupçonnées. Il avance en *terra incognita*, et même s'il a réalisé un travail prospectif très complet en amont du lancement de son programme, il peut s'attendre de façon certaine à rencontrer beaucoup d'imprévus et à se heurter à de multiples obstacles, petits et grands. Cela implique des coûts logistiques souvent très importants et des délais de production parfois conséquents. De ce point de vue, le temps de l'évaluation n'est pas toujours conforme avec le temps de la décision publique.

Ces obstacles seront d'ordre juridique ou réglementaire (en relation avec la multiplicité des acteurs et des contrats ou conventions qui vont les relier), d'ordre logistique (les conditions d'accueil des personnes vont soulever de nombreux problèmes de disponibilité, de sécurité des locaux, …), liés aux chaînes de traitement des données de gestion et aux systèmes d'information (qui peuvent s'avérer incomplets ou inadaptés), ou encore d'ordre organisationnel (la mise en œuvre de l'expérimentation suppose non seulement l'adhésion du porteur de projet mais également celle de tous les acteurs qui sont amenés à mettre en pratique

l'expérimentation),. S'ils n'adhèrent pas, ces acteurs peuvent mettre en œuvre des stratégies visant à contourner le protocole et qui pourront parfois conduire à attribuer tout ou partie du traitement, voire un traitement de substitution, au groupe témoin, ce qui fausserait l'évaluation. L'évaluateur va lui aussi être nécessairement confronté à un grand nombre d'imprévus et d'impondérables. S'il pense évaluer une expérimentation, en réalité il expérimente une évaluation.

Dans ce contexte, il faut souligner que le fait de vouloir mener à bien une évaluation rigoureuse des effets d'un programme expérimental ajoute des complications et des incertitudes dans le déroulement d'une expérimentation. Il en résulte des coûts de mise en œuvre, de suivi et de coordination pour les parties prenantes. Ces coûts ne sont pas forcément déterminants dans le choix stratégique du protocole d'évaluation mais ils doivent néanmoins être pris en considération par l'expérimentateur et son évaluateur.

L'évaluation quantitative requiert la constitution de bases de données ayant à la fois une dimension individuelle et temporelle. Ces bases agrègent deux types de sources statistiques. D'un côté, on mobilise les bordereaux de suivi de l'expérimentation, qui consistent dans des fiches individuelles qui donnent la date d'entrée dans le programme, le type de parcours suivi dans le programme, la date de sortie, et d'autres indications sur le déroulement du programme et la situation des personnes avant et après le programme, ainsi que des fiches collectives, sur les différents opérateurs et leurs prestations. D'un autre côté, on mobilise des fichiers administratifs de gestion qui permettent de suivre les personnes du point de vue de l'expérimentateur et des organismes gestionnaires (tels que Pôle Emploi ou la CAF). Toutes ces sources statistiques doivent être appariées de façon à disposer d'une information complète sur les trajectoires et les caractéristiques des personnes avant, pendant et après l'expérimentation, qu'elles appartiennent au groupe test ou au groupe témoin. Puis, des traitements statistiques et économétriques sont réalisés sur ces bases de données.

La production d'une telle base de données est un travail statistique lourd. S'agissant de données individuelles, elle peut poser en outre des questions de secret statistique et de respect du cadre législatif sur l'informatique et les libertés. L'expérimentateur est le producteur et le gestionnaire des fichiers de gestion. Il organise également, de façon directe ou indirecte, la production des données de suivis tout au long de l'expérimentation, avec l'appui technique de l'évaluateur. Ce dernier est destinataire de la base donnée. Afin de respecter les textes en vigueur relatifs au secret statistique et à la confidentialité des données individuelles,

l'évaluateur ne doit disposer ni même manipuler aucune information directement ou indirectement nominative. La base doit donc être complètement anonyme et les appariements de fichiers doivent être effectués en amont, par l'expérimentateur. Il s'agit là encore d'un surcroît de travail non négligeable. Cet investissement pourra cependant être amorti à la fois à des fins d'évaluation interne et externe, à des fins descriptive et explicative. La constitution de ce système d'observation pourra par ailleurs permettre d'améliorer la gestion même du dispositif.

Conclusion

Comment évaluer une méthode d'évaluation ? Sa capacité à produire de nouvelles connaissances est certainement un critère à considérer. De ce point de vue de productivité heuristique, l'évaluation aléatoire a largement fait ses preuves. Comme le soulignent Banerjee et Duflo (2008), l'un des principaux apports de ces méthodes expérimentales est qu'elles permettent de couvrir des terrains ignorés des approches non expérimentales. Ce faisant, elles n'ont pas vocation à se substituer aux autres méthodes d'évaluation, d'autant qu'elles sont inapplicables dans bien des cas (par exemple pour des politiques macroéconomiques, pour évaluer des traitements vitaux, pour des mesures ponctuelles et instables dans le temps, …). Elles ont donc bien vocation à compléter la boîte à outil de l'évaluateur de programmes sociaux. Ce faisant, elles étendent la responsabilité du chercheur à la production des données et parfois même, à celle des politiques publiques qui sont mises en œuvre. Elles étendent aussi nos champs de recherche dans des domaines inexplorés. Si ces nouvelles méthodes d'évaluations ne sont pas mises en oeuvre, la majorité des programmes sociaux locaux et expérimentaux ne sera tout simplement pas évaluée.

Références

Angrist J. et A. Krueger, (1999), « Empirical Strategies in Labor Economics », in Handbook of Labor Economics, vol 3A, Ashenfelter O. et Card D. (eds.), North Holland, Amsterdam, p. 1277-1366.

Banerjee A. et E. Duflo (2008). «The Experimental Approach to Development Economics ». Mimeo MIT J PAL.

Behaghel L., Crépon B. et Gurgand M. (2009). « Evaluation d'impact de l'accompagnement des demandeurs d'emploi par les opérateurs privés de placement et le programme Cap vers l'entreprise », rapport final, miméo, septembre.

Blundell R. et Costa Dias M. (2000). « Evaluation Methods for Non-Experimental Data » ; *Fiscal Studies,* 21(4), 427–468

Bruhn M., McKenzie D. (2009). "In pursuit of balance : randomization in practice in development field experiments", *American Economic Journal: Applied Economics*, American Economic Association, vol. 1(4), pages 200-232, October

Burtless G., (1995). «The Case for Randomized Field Trials in Economic and Policy Research», *Journal of Economic Perspectives*, Vol 9, n°2, pp 63-84.

Deaton A. (2009). "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic Development", mimeo Princeton University, janvier.

Dehejia R. et Wahba S., (1999), « Causal Effects in Non experimental Studies: re-evaluating the evaluation of training programs », *Journal of the American Statistical Association*, 94, p. 1053- 1062.

Duflo E. , Glennerster R. Kremer M. (2006). « Using Randomization in Development Economics Research: A Toolkit », *MIT Department of Economics Working Paper, No. 06-36*

Dufflo E. (2009). « Expériences, Science et Lutte contre la pauvreté ». Leçon inaugurale au collège de France, Chaire internationale "Savoirs contre pauvreté", 8 janvier 2009.

Dufflo E. (2010-a). *Le développement humain. Lutter contre la pauvreté (I)*, Le Seuil / République des idées, Paris. 104 p.

Dufflo E. (2010-b). *La politique de l'autonomie. Lutter contre la pauvreté (II)*, Le Seuil / République des idées, Paris. 104 p.

Fougère D. (2000). « Expérimenter pour évaluer les politiques d'aide à l'emploi : les exemples anglo-saxons et nord-européens », *Revue Française des Affaires sociales*, n°1, janvier-mars.

Goujard A. et Y. L'Horty (2010). « La définition des zones témoins du Revenu de Solidarité Active », *Revue Française des Affaires Sociales*, n°1-2, janvier-juin, 64 ème année.

Gouriéroux C. (1981). *Théorie des sondages*. Economica.

Gurley T. et Bruce D. (2005), "The effects of car access on employment outcomes for welfare recipients", *Journal of Urban Economic*s, 58 (2), pp. 250-272.

Heckman J. (1992). "Randomization and social policy evaluation," in Evaluating Welfare and Training Programs, editors Charles Manski and I. Garfinkel. Cambridge, MA: Harvard University Press. (also available as NBER Technical Working Paper No.107, 1991).

Heckman J.J., LaLonde R.J. et Smith J.A., (2000), "The Economics and Econometrics of Active Labor Market Programs", in Handbook of Labor Economics, vol 3A, Ashenfelter O. et Card D. (eds.), North Holland, Amsterdam, p. 1865-2097.

Heckman, J. (2000). "Microdata, Heterogeneity and the Evaluation of Public Policy", Nobel Prize Lecture, December 8.

Imbens, Guido and Jeffrey M. Wooldridge (2008). "Recent Developments in the Econometrics of Program Evaluation," MIMEO, Harvard University (forthcoming in *Journal of Economic Literature*).

Kremer M. (2003). "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons" *American Economic Review*, 93(2), pp 102-106.

Levitt, S. D. et List, J. A., (2008). "Field Experiments in Economics: The Past, the Present, and the Future" (September 2008). NBER Working Paper No. W14356.

L'Horty Y., Petit P. (2009). *Guide méthodologique pour l'évaluation des expérimentations sociales*. Miméo, http://www.experimentationsociale.fr

Lopez A. Rouaud. P. (2010). "Expérimentations sociales: lorsque l'assignation aléatoire porte sur des groupes, Usage de la macro CUBE", *Relief* n°30, Echanges du Cereq, mai, pp 157-174..

Meyer B., (1995), « Natural and Quasi-Experiments in Economics », *Journal of Business and Economic Statistics*, 13(2), p. 151-62.

Miguel E. et M. Kremer (2004) « Worms: Education and Health Externalities in Kenya » *Econometrica*, Vol. 72, No. 1 (January), 159–217

Ong P. (2002), "Car Ownership and Welfare-to-Work", *Journal of Policy Analysis and Management*, 21 (2), pp. 239-252.

Raphael S. et Rice L. (2002), "Car Ownership, Employment and Earnings", *Journal of Urban Economics*, 52, pp. 109-130.

Rubin D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies", *Journal of Educational Psychology*, (66): p. 688–701.

Rodrik D. (2008). "The New Development Economics: We Shall Experiment, but How Shall We Learn?" *HKS Faculty Research Working Paper Series*, n° 08-055