# Analyse d'impact : l'apport des évaluations aléatoires

**William Parienté**[1]

**Cet article expose de manière synthétique l'apport des évaluations aléatoires dans l'analyse de l'impact des programmes sociaux et de développement. Ces évaluations reposent sur l'assignation aléatoire d'un groupe recevant le programme (traitement) et d'un groupe ne le recevant pas (contrôle). La comparaison de ces groupes après intervention permet d'obtenir une mesure non biaisée de l'impact du programme. La méthode des évaluations aléatoires surmonte de ce fait un nombre important de limites des évaluations non-expérimentales. Elle partage aussi certaines difficultés inhérentes à toute évaluation telle que la possibilité de généraliser les résultats. La portée des évaluations aléatoires ne se résume pas seulement à mesurer les effets d'un programme, elles peuvent aussi être utilisées pour comparer différentes modalités d'une intervention (et ainsi identifier la plus efficace), tester des innovations introduites dans un programme ou encore analyser des hypothèses de la théorie économique. Si le principe des évaluations aléatoires est simple, la conception des dispositifs expérimentaux est déterminante pour permettre de mesurer correctement les effets d'un programme.**

## Introduction

Comment mesurer l'efficacité d'un programme scolaire, d'un programme de prévention santé, d'une intervention visant à améliorer l'accès au crédit des populations pauvres, ou encore d'une politique sur le marché du travail ?

Lors de la mise en œuvre de politiques sociales ou de développement, la question de leur efficacité est systématiquement posée**.** Cependant, plusieurs types de réponses peuvent être envisagés :

- Une *évaluation des besoins des populations* concernées revient à analyser si les besoins des populations ciblées ont été effectivement pris en compte et si la population bénéficiant du programme a été convenablement identifiée.
- Une *évaluation du processus* cherche à mesurer l'efficacité d'un programme en vérifiant la nature du processus de sa mise en œuvre, si les services sont effectivement délivrés, s'ils sont de qualité, si les bénéficiaires en sont satisfaits etc.
- Enfin, une *évaluation d'impact* consiste à évaluer l'impact direct du programme sur les bénéficiaires.

Cet article porte sur ce dernier type d'évaluation. S'il existe différentes méthodologies d'évaluation d'impact, cet article expose de manière synthétique essentiellement les évaluations aléatoires, qui ont connu un engouement croissant au cours des dix dernières années[2].

Dans un contexte où l'efficacité de l'aide et des politiques publiques est régulièrement remise en question, le développement des évaluations aléatoires est né en partie du constat selon lequel il existe un nombre limité de *preuves empiriques rigoureuses* sur ce qui « marche » et « ne marche pas » dans les programmes sociaux ou de développement. En effet, l'efficacité des programmes est rarement évaluée de façon

[1] *JPAL – PSE/Ecole d'Economie de Paris. Email : william.pariente@parisschoolofeconomics.eu.*

[2] *Les lecteurs désireux d'aller plus loin dans la compréhension de la méthode d'évaluation aléatoire peuvent se reporter aux travaux de J-PAL (Jameel Poverty Action Lab, www.povertyactionlab.com) ou aux travaux mis en ligne par l'initiative 3ie « International Initiative for Impact Evaluation » au www.3ieimpact.org.*

rigoureuse. Les évaluations aléatoires, lorsqu'elles sont correctement mises en œuvre, permettent d'obtenir des mesures de l'impact non biaisées et surmontent de ce fait un nombre important de problèmes internes aux évaluations non-expérimentales.

A l'origine, la plupart des évaluations aléatoires ont été menées dans les pays industrialisés, aux États-Unis et en Europe du Nord à partir des années 1960, sur des politiques de formation et d'emploi, de santé ou de logement. De nombreux programmes ont été évalués aux Etats-Unis par l'organisation MDRC[1] et ont eu un impact direct sur les politiques publiques (Burtless, 1995). Dorénavant, un nombre important d'évaluations aléatoires sont menées également dans les pays en développement en collaboration avec les organisations non gouvernementales (ONG), gouvernements ou organisations internationales. Ces évaluations couvrent aujourd'hui de nombreux domaines : éducation, santé, adoption de nouvelles technologies, accès au crédit, lutte contre la corruption, politiques sur le marché du travail etc.

Cet article est organisé en plusieurs sections : la première section introduit le principe de l'évaluation d'impact et l'apport de la méthode aléatoire, la seconde analyse la portée de ces évaluations sur la mise en place des politiques publiques mais aussi, par certains aspects, sur la théorie économique. La troisième énumère certaines limites des évaluations d'impact en général et les limites spécifiques aux évaluations aléatoires. La quatrième aborde la mise en œuvre pratique des évaluations, les questions de faisabilité opérationnelle et de validité statistique. La dernière introduit des éléments de présentation de résultats des évaluations. Ensuite vient la conclusion.

## Principe de l'évaluation d'impact

Le modèle statistique de base des évaluations est le modèle *causal* de Rubin (1974). Les unités sont par exemple des individus, repérés par un indice *i*. Le traitement T est administré de manière binaire : $T_i = 1$ signifie que l'individu i appartient au groupe qui reçoit le traitement, $T_i = 0$ signifie qu'il ne reçoit pas le traitement.

On s'intéresse à des variables de résultat, sur lesquelles le traitement est supposé avoir un impact. Le modèle causal de Rubin considère que pour une variable de résultat donnée il y a en fait deux variables dites d'*outputs potentiels* ou *latents*, correspondant à ce que serait la situation d'un individu sous chacune des alternatives, *i.e.* si l'individu bénéficie du traitement $y_i(1)$ et s'il n'en bénéficie pas, $y_i(0)$.

Afin de mesurer exactement l'impact d'un programme, il est nécessaire de connaître la situation potentielle des bénéficiaires du programme si jamais ceux-ci n'y avaient pas participé. L'effet du traitement se résume donc à $E(y_i(1)) - E(y_i(0))$.

Cet effet est donc *individuel* et peut être, tel que défini, *hétérogène* au sein de la population. Il est aussi *inobservable*, puisque pour chaque individu, soit on observe $y_i(1)$ lorsqu'il dispose du programme, $T_i = 1$, soit on observe $y_i(0)$ lorsqu'il n'en dispose pas $T_i = 0$, mais jamais simultanément les deux.

Dans une population donnée, on dispose en fait d'observations avec $T_i = 1$ et des observations avec $T_i = 0$ mais elles ne concernent pas les mêmes individus :

$$
\begin{aligned}
& E(y_i|T=1) - E(y_i|T=0) \\
&= E(y_i(1)|T=1) - E(y_i(0)|T=0) \qquad (1) \\
&= E(y_i(1) - y_i(0)|T=1) + E(y_i(0)|T=1) - E(y_i(0)|T=0) \qquad (2)
\end{aligned}
$$

L'effet moyen du traitement est compris dans la première composante de (2) $E(y_i(1) - y_i(0)|T=1)$ alors que la deuxième composante $E(y_i(0)|T=1) - E(y_i(0)|T=0)$ comprend le biais de sélectivité ou l'effet population.

L'objectif d'une évaluation d'impact est de réduire le plus possible ce biais de sélectivité en comparant le groupe de bénéficiaires du programme à un groupe de non bénéficiaires ayant exactement les mêmes caractéristiques exceptée la participation au programme.

En pratique, il est très difficile de trouver un tel groupe. Si certaines caractéristiques sont facilement mesurables, comme par exemple, le degré de richesse ou le niveau de scolarisation des individus, il existe, en revanche, des caractéristiques spécifiques souvent inobservables qu'il est très difficile de prendre en compte comme par exemple le niveau de motivation, l'accès à l'information sur l'existence d'un projet, les capacités des individus, etc. En général, ceux qui participent à un

[1] *Manpower Demonstration Research Corporation.*

programme sont justement les plus motivés ou les plus informés.

La *randomisation* de l'assignation du traitement permet d'éliminer cet effet de sélection et joue donc un rôle central dans la possibilité d'identifier de manière rigoureuse l'effet du programme. Dans le cas de l'assignation aléatoire du traitement entre un groupe *traitement* et un groupe *contrôle*, on a :

$$E(y_i(0)) \perp T$$

Et

$$E(y_i(0)|T=1) - E(y_i(0)|T=0) = 0$$

La sélection aléatoire, assure, si l'échantillon est suffisamment grand, que les caractéristiques observables mais aussi inobservables des individus sont similaires dans les groupes traitement et contrôle, avec par exemple le même nombre d'individus riches et pauvres, le même niveau de scolarisation mais aussi le même degré de motivation ou le même niveau d'information.

L'effet population ou le biais de sélectivité est, par construction, égal à zéro. De ce fait, toutes différences identifiées entre les deux groupes après le programme, peuvent être entièrement attribuées au programme.

## Evaluations non expérimentales

Les évaluations aléatoires permettent donc de résoudre le problème de biais de sélectivité. D'autres méthodes, non aléatoires, dites « *quasi-expérimentales* », tentent de résoudre le problème de biais de sélectivité sous une série d'hypothèses permettant de reconstruire les conditions d'un cadre expérimental.

Une première méthode consiste à reconstituer un groupe de contrôle à partir de variables observables au sein d'un ensemble d'individus n'ayant pas reçu le traitement. L'objectif est ainsi de supprimer le biais de sélectivité en comparant deux populations ayant des caractéristiques observables similaires. Pour ces méthodes dites *méthodes d'appariement*, il s'agit de trouver pour chaque individu dans le groupe traité, un individu ayant les mêmes caractéristiques observables dans le groupe non traité. Ce type d'appariement est rendu difficile en pratique lorsqu'il est effectué sur un grand nombre de variables ou des variables continues car il peut être compliqué de trouver dans le groupe d'individus non traités, un nombre suffisant d'individus ayant exactement les mêmes caractéristiques que les individus traités. Des méthodes d'appariement proposent d'utiliser un *score de propension* pour permettre d'apparier sur un plus grand nombre de variables, et notamment des variables continues.

Ces méthodes prennent en compte les différences d'output imputables potentiellement à des différences entre les groupes traitement et contrôle sur un grand nombre de variables observées. Elles parviennent cependant difficilement à supprimer entièrement le biais de sélectivité car certaines variables inobservables (telles que, à nouveau, la motivation des individus, l'accès à l'information, etc.) sont omises et ne sont pas, par essence, inclues dans la construction de l'appariement. Dans ce cas, la méthode d'appariement atténue probablement le biais de sélectivité sur la base des observables, mais n'a aucune incidence sur le biais induit par les variables omises.

Ensuite, les estimations en *double–différence* (souvent combinées avec les méthodes d'appariement) utilisent les informations avant intervention des groupes traitement et contrôle pour contrôler les différences préexistantes entre les deux groupes. Cependant, la validité des estimations repose sur le parallélisme de l'évolution entre les deux groupes. En effet si les deux groupes devaient évoluer de manière très différente (par exemple, si les personnes qui participent au programme avaient une propension à augmenter leurs revenus plus rapidement que ceux qui ne participent pas), le biais de sélectivité resterait important car cet écart du aux inobservables est attribué au programme.

Enfin, d'autres évaluations utilisent la *discontinuité* dans la règle de mise en œuvre d'un programme, souvent établie de manière arbitraire, pour analyser son effet. Par exemple, Buddelmeyer et Skoufias (2003) évaluent l'impact du programme d'allocations conditionnelles *Progresa* au Mexique sur l'absentéisme scolaire et le travail des enfants en utilisant la discontinuité dans les critères d'éligibilité[2]. L'idée sous jacente de ce type d'évaluation est d'estimer l'effet en utilisant les individus juste en dessous du seuil d'éligibilité comme contrôle par rapport à ceux qui sont juste au dessus. Cette méthode est comparable aux évaluations aléatoires si ce n'est que l'impact est évalué uniquement sur une population spécifique (juste en dessus et en dessous du seuil). En pratique, il peut s'avérer difficile de disposer de règles exogènes permettant l'utilisation de ce type de méthodes[3].

---

*[2] Le programme Progresa a initialement été évalué de manière aléatoire. Les auteurs montrent que pour 10 indicateurs sur 12, les résultats avec la méthode régression en discontinuité sont similaires aux résultats de l'évaluation aléatoire.*

*[3] Notamment dans les pays en développement où les règles d'éligibilité peuvent ne pas être totalement respectées (Duflo et al., 2007). D'autre part, ce type de méthode n'est possible que lorsqu'il existe un nombre*

Les méthodes d'appariement et/ou double différence peuvent permettre d'atténuer significativement le biais de sélectivité[4] sans pour autant l'éliminer totalement et reposent sur un nombre d'hypothèses qui peuvent être plus ou moins convaincantes alors que d'autres, telles que les régressions en discontinuité, peuvent dans certains cas donner des résultats non biaisés.

## Portée des évaluations aléatoires

La portée des évaluations aléatoires se situe à plusieurs niveaux. En général, les évaluations sont mises en place pour tester l'impact d'une intervention sur les bénéficiaires. Il peut s'agir d'un programme pilote (par exemple gouvernemental) qui sera généralisé si les résultats sont probants. Dans d'autres cas les évaluations sont effectuées sur des interventions plus spécifiques au niveau d'un programme particulier souvent mis en œuvre par des ONG. Certaines évaluations cherchent à analyser plusieurs interventions ou modalités d'intervention afin d'identifier les meilleurs moyens pour atteindre un objectif donné. Enfin d'autres évaluations ne sont pas directement liées à l'impact d'un programme mais cherchent plutôt à analyser l'effet d'innovations sur le fonctionnement d'un programme ou à répondre à une question théorique spécifique.

### Projets pilotes

Les évaluations aléatoires peuvent intervenir dans le cadre d'un projet pilote visant à être généralisé si les résultats sont concluants. Ces évaluations sont en général réalisées sur des échantillons représentatifs de la population cible.

L'exemple le plus connu dans les pays en développement est celui du programme *Progresa* au Mexique. Le gouvernement mexicain a développé un programme d'allocations aux ménages représentant environ un tiers de leurs revenus. Ces allocations étaient conditionnées à l'assiduité des enfants à l'école et à la participation du ménage à des sessions de prévention santé. Lorsque le programme a été lancé en 1998, le gouvernement ne pouvait pas atteindre l'ensemble des localités éligibles du fait de contraintes budgétaires. D'autre part, l'équipe en place était consciente que pour assurer la pérennité du projet, il fallait être capable de démontrer son efficacité pour qu'il ne soit pas abandonné en cas d'alternance après les élections. Le gouvernement a décidé, en collaboration avec des universitaires de l'IFPRI[5], de lancer un programme pilote sur un nombre restreint de localités en milieu rural choisies aléatoirement parmi un ensemble de localités éligibles (320 sur 506 et 24 000 ménages). Les résultats de l'évaluation sont positifs sur un grand nombre d'indicateurs. Le programme accroît notamment la participation des enfants à l'école, réduit la prévalence de maladie chez les enfants, diminue l'offre de travail des garçons, et permet aux parents de perdre moins de jours de travail à cause des maladies (Gertler and Boyce, 2001). Grâce à ces résultats positifs, le projet a été généralisé au Mexique pour atteindre plus de 5 millions de famille en 2005 et a été étendu au secteur urbain. D'autres pays d'Amérique latine, le Nicaragua, le Honduras ou l'Argentine, par exemple, ont adopté des programmes similaires.

La difficulté à généraliser à une échelle plus large les résultats des évaluations aléatoires le plus souvent effectuées dans un contexte particulier est souvent citée comme une limite des évaluations. Ce point est abordé en troisième partie.

### Évaluations d'interventions spécifiques

Les évaluations aléatoires sont également utilisées pour analyser les effets d'une intervention spécifique sur les bénéficiaires. La portée des résultats de ces évaluations va cependant souvent au-delà de l'impact du programme lui-même.

Dans le domaine de l'éducation, plusieurs évaluations aléatoires (Miguel et Kremer, 2004 et Miguel et Kremer, 2007), menées en collaboration avec l'ONG ICS Africa[6] (International Child Support - Africa) montrent l'impact très significatif sur l'assiduité scolaire des programmes de déparasitage des vers intestinaux des enfants. Dans le contexte du Kenya, le déparasitage en masse (où chaque enfant est traité) est le moyen le plus efficace et le moins coûteux pour augmenter l'assiduité scolaire des enfants. En effet, le déparasitage accroît la participation scolaire de 7 % et permet de réduire de 25 % le taux d'absentéisme à l'école. En comparaison avec d'autres programmes qui visent à améliorer l'assiduité des enfants dans les pays en développement, le

*suffisant d'individus juste en dessous et juste au dessus du seuil d'éligibilité.*

*[4] Dans la décomposition de l'effet global, Heckman et al. (1997) montrent qu'avec les méthodes quasi-expérimentales, l'effet population est environ équivalent à l'effet traitement. D'autre part, Lalonde (1986) montre que les estimations de l'effet de plusieurs programmes avec des méthodes expérimentales et quasi-expérimentales diffèrent de manière significative. Glazerman, Levy and Myers (2003) arrivent à des conclusions similaires lorsqu'ils comparent les résultats d'évaluations quasi-expérimentales et expérimentales sur des programmes d'emploi et de formation.*

*[5] International Food Policy Research Institute.*

*[6] ICS est spécialisée dans des projets d'éducation, de santé et de soutien à des activités génératrices de revenus.*

déparasitage serait environ 20 fois plus efficace que d'embaucher des enseignants supplémentaires. Pour les pays en développement où la prévalence des vers intestinaux chez les enfants est élevée[7], les programmes de déparasitage pourraient donc s'avérer être l'un des moyens les plus performants tout en étant l'un des moins onéreux pour accroître la participation scolaire.

Dans certains cas, des programmes sont évalués dans différents contextes pour donner des informations pertinentes en cas de généralisation. Banerjee *et al.* (2007) ont évalué l'impact d'un programme de rattrapage scolaire simultanément dans toutes les écoles de deux villes très différentes en Inde (Vadodara et Mumbai). Après deux ans, l'augmentation des tests standardisés des élèves était en moyenne de 0,39 écarts types et l'impact était plus fort pour les élèves les plus faibles. L'obtention de résultats positifs dans deux contextes différents pourrait rendre crédible la généralisation de l'intervention.

La multiplication d'évaluations sur des programmes similaires dans des contextes différents permet d'améliorer la connaissance générale de l'impact de certaines interventions. Par exemple, de plus en plus de programmes de microfinance sont en train d'être évalués (aucun résultat n'est encore paru à ce jour). Ces évaluations ont lieu notamment en milieu urbain en Inde (Banerjee et Duflo), en milieu rural au Maroc (Crépon *et al.*) ou au Pérou (Karlan et Valdivia). Elles sont d'autant plus importantes que peu d'évaluations de la microfinance[8] ont donné jusqu'à présent des preuves empiriques vraiment rigoureuses de l'impact. Or, si la microfinance constitue un succès indéniable au regard du nombre de personnes qui ont aujourd'hui accès au micro-crédit, la question de son impact sur les conditions de vie et sur la pauvreté n'en demeure pas réglée. Les résultats des évaluations en cours permettront de contribuer à l'amélioration de la connaissance sur l'impact de la microfinance.

## Modalités d'intervention

Les évaluations peuvent tester plusieurs modalités d'intervention pour sélectionner celles qui sont les plus efficaces en terme d'impact ou de moyens pour toucher une population donnée.

Dans le domaine de la prévention, Duflo *et al.*, (2006) comparent l'impact de trois modes de prévention au VIH/SIDA : la formation des professeurs, l'encouragement de débats entre étudiants sur la protection, et la réduction du coût de l'éducation. L'impact est estimé sur le nombre de grossesses liées à un rapport sexuel non protégé, ainsi que sur des mesures de la connaissance et du comportement vis-à-vis du VIH/SIDA. Après deux ans, les filles ayant suivi les cours de professeurs formés ont plus de chance d'être mariées au moment de leur grossesse mais le programme a eu très peu d'autres impacts sur la connaissance, le comportement ou simplement le nombre de grossesses chez les adolescentes. Les débats sur la protection ont accru la connaissance vis-à-vis du VIH/SIDA, ainsi que la déclaration d'utilisation de préservatifs mais n'ont pas eu d'impact sur l'activité sexuelle. La réduction du coût de l'éducation en offrant des uniformes scolaires a servi à diminuer les décrochages scolaires, les mariages et les grossesses chez les adolescentes. Les trois modes de prévention ont donc des impacts positifs mais sur différentes variables d'output étudiées. En terme de protection, il semble que les débats sur l'utilisation des préservatifs soient les plus efficaces. Enfin, une analyse ultérieure sur l'effet des programmes sur la prévalence du VIH/SIDA et sur le coût de chaque intervention permettrait d'identifier celle qui est la plus efficace.

Dans un autre domaine, Olken (2007) s'intéresse à différents moyens de réduire la corruption en Indonésie en testant plusieurs systèmes de contrôle des projets de construction de routes. Les deux systèmes envisagés sont l'augmentation d'audits du gouvernement (avec des changements de probabilité d'audit allant de 4 à 100 %) et deux méthodes de renforcement de la participation de la communauté dans la gestion du suivi des projets. La corruption est mesurée par la différence entre le coût de la construction effective (mesurée par des ingénieurs et des enquêtes ex-post) et le coût apparaissant dans le budget des villages.

Les résultats montrent que l'augmentation de la probabilité de contrôles externes (audits du gouvernement) réduit significativement le montant des fonds manquants au projet (de 27,7 à 19,2 points de pourcentage). En revanche, l'implication de la communauté dans la gestion du suivi des projets n'a un effet que sur certaines formes de corruption (sur les dépenses liées aux ressources humaines), mais n'a pas d'impact général. L'efficacité des encouragements à contrôler des projets d'infrastructure au sein de la communauté demeure relativement faible. En termes de décisions de politiques publiques, il apparaît, dans ce contexte, que pour lutter contre la corruption il est plus efficace d'augmenter les contrôles externes (par une autorité gouvernementale) que d'impliquer plus les membres de la communauté.

*[7] Environ 400 millions d'enfants en âge d'aller à l'école seraient affectés par les vers intestinaux (WHO, 2004).*

*[8] L'évaluation de Pitt et Kandker (1998) du micro-crédit au Bangladesh est un exemple d'évaluation quasi-expérimentale rigoureuse même si ses résultats ont été contestés (Morduch, 1998).*

## Innovations du processus des interventions

Certaines évaluations testent des innovations du processus des interventions ou des politiques. Par exemple, en microfinance, Giné et Karlan (2006) évaluent l'impact d'un changement de produit d'une institution de microfinance aux Philippines qui octroyait essentiellement des crédits de groupe (garantie solidaire). En effet, les taux de remboursement très élevés des institutions de micro-crédit dans le monde sont attribués, en partie, à la méthode de garantie solidaire. L'évaluation a consisté à changer le type de crédit offert dans la moitié des agences de l'institution, en passant du crédit de groupe au crédit individuel. Les résultats montrent que non seulement, le changement de type de crédit n'a pas eu d'impact sur le taux de remboursement, mais a eu un impact positif sur la rétention des clients et sur la participation de nouveaux clients.

## Tests d'hypothèses de la théorie économique

Enfin, les évaluations aléatoires permettent de répondre à des questions théoriques qui ont des implications importantes en terme de politique publique sans être uniquement limitées à l'évaluation d'un programme.

Par exemple, une question récurrente dans l'offre de services ou de biens de base aux plus pauvres est l'élasticité de la demande par rapport au prix et l'impact du prix sur l'utilisation de ces biens.

Un consensus existe sur l'importance de fournir aux populations pauvres des services ou des biens liés à la santé de manière subventionnée. En revanche il y a débat sur la nécessité de faire participer financièrement ces populations, même de manière très restreinte.

La participation financière permettrait d'augmenter l'intensité de l'utilisation du bien en question et de sélectionner ceux qui en ont le plus besoin (et qui sont prêts à payer pour l'obtenir). Un travail sur le prix des moustiquaires au Kenya (Cohen et Dupas, 2007) donne des indications importantes en terme d'efficacité des mécanismes des politiques publiques et sur des hypothèses théoriques. Dupas et Cohen (2007) mettent en place un dispositif original de distribution de moustiquaires dans 20 cliniques au Kenya pour les femmes enceintes. Ces cliniques ont été divisées en 2 groupes ; un groupe contrôle et un groupe où les moustiquaires étaient distribuées à 4 prix différents variant de 0 à 60 centimes de dollar. Le choix des groupes et des prix étant effectué de manière aléatoire.

Les résultats montrent, d'une part, que la demande pour les moustiquaires diminue fortement lorsqu'elles sont vendues même à un prix relativement faible (inférieur au prix vendu par une ONG de la même région). Pour 100 femmes prenant la moustiquaire lorsqu'elle est gratuite, seulement 25 l'achèteraient au prix proposé.

Les résultats montrent également que l'intensité de l'utilisation des moustiquaires ne varie pas avec le prix. L'effet psychologique du prix sur l'utilisation est alors mineur.

Finalement, ceux qui consentent à payer un prix ne semblent pas non plus être ceux qui en ont le plus besoin (le besoin étant mesuré par le taux d'anémie qui est un indicateur important de la malaria). Ceux qui consentent à payer ne sont pas plus malades que ceux qui acceptent la moustiquaire gratuitement. Il est probable que les plus malades aient en effet une volonté à payer plus forte mais aussi qu'ils soient les plus pauvres.

Ces résultats remettent en question, pour ce bien spécifique, la politique de participation financière des ménages pauvres. Ils montrent aussi qu'une proportion importante de la population vulnérable qui a besoin de moustiquaires risque de ne pas être servie par le programme.

Finalement, en termes d'impact sur la santé, il est plus efficace de fournir les moustiquaires gratuitement.

Dans un autre domaine, Karlan et Zinman (2005) cherchent à tester l'intensité des facteurs responsables de l'asymétrie d'information sur le marché du crédit. Le dispositif d'évaluation consiste à proposer, aléatoirement, plusieurs offres de contrats de crédit avec différents taux d'intérêt (bas et élevés) et, une fois acceptées, à varier le taux d'intérêt effectivement proposé et donner aléatoirement un encouragement spécifique (un taux d'intérêt bas) en cas de remboursement satisfaisant. L'intensité de la sélection adverse (*ex ante*) est mesurée en comparant le taux de remboursement des clients ayant accepté des taux d'intérêt différents mais ayant effectivement reçu le même taux (le plus bas). L'intensité de l'aléa moral est quant à elle mesurée en comparant le taux de remboursement des clients ayant accepté la même offre au même taux mais qui ont effectivement reçu des contrats avec des taux différents et en comparant les performances de remboursements des clients qui ont reçu et n'ont pas reçu l'encouragement. Les résultats montrent qu'il existe un aléa moral significatif mais seulement une faible présence de sélection adverse.

Qu'il s'agisse du pilote d'une politique publique, d'un programme local, d'une innovation ou du test d'une hypothèse théorique, l'ensemble des résultats des évaluations aléatoires permettent d'accumuler du savoir sur les effets de certains projets ou évaluations. Ils constituent, en ce sens, un bien public qui peut être utile pour les décideurs politiques ou les praticiens dans la mise en œuvre de leurs programmes.

Enfin, l'analyse des résultats des évaluations permet de comparer le rapport efficacité – coût des programmes si ceux-ci évoluent dans des contextes similaires et ont des objectifs propres. Par exemple, des programmes de déparasitage au Kenya s'avèrent être beaucoup plus efficaces sur l'assiduité scolaire et moins coûteux que d'autres programmes (enseignants supplémentaires, uniformes, etc.). Certaines analyses cherchent également à comparer directement les coûts et les bénéfices d'un programme pour la société. Il peut cependant s'avérer compliqué de comparer le coût monétaire d'un programme avec des bénéfices qui ne peuvent pas toujours être valorisés en termes monétaires.

## Limites des évaluations aléatoires

Deux grands types de questions se posent quant à la validité des évaluations. Le premier concerne la *validité interne :* il s'agit de vérifier si les effets mesurés sont bien les effets du programme, si les groupes traitement et contrôle sont comparables, etc. Le second type de question concerne la *validité externe*, liée au caractère généralisable des résultats de l'étude dans d'autres contextes ou à la représentativité de l'échantillon par rapport à la population cible.

La validité interne est maximale dans le cas des évaluations aléatoires (et dans certains contextes pour les méthodes de *régression en discontinuité*). Les évaluations aléatoires partagent cependant certaines limites avec l'ensemble des évaluations et comportent dans certains cas des limites plus spécifiques.

En termes de validité interne, il est possible que, malgré la randomisation, la comparabilité des groupes traitement et contrôle puisse être compromise. Il est par exemple possible qu'en raison de leur participation à une expérimentation, les individus modifient leur comportement. Ces modifications de comportement au sein du groupe traitement sont appelées *effets Hawthorne* tandis que ceux au sein du groupe contrôle sont appelés *effets John Henry*. Les deux groupes peuvent être incités à modifier leur comportement car ils se savent observés mais aussi parce qu'ils sont gagnants ou perdants. Ces modifications de comportement dues au cadre expérimental posent des problèmes de validité interne et externe. Il y a différentes manières de limiter ces effets voire de les mesurer selon le dispositif d'évaluation choisi.

Un autre problème potentiel lié à la comparabilité des deux groupes que l'on peut observer ex post est l'attrition. L'attrition correspond à l'échec de collecter des données sur les variables d'intérêt pour certains individus qui faisaient partie de l'échantillon original. L'attrition aléatoire réduira seulement la puissance statistique de l'étude mais si l'attrition n'est pas indépendante du traitement, les résultats peuvent être biaisés (dans certains cas, il est plus facile de retrouver les individus du groupe traitement que ceux du groupe contrôle).

D'autres problèmes tels que l'adhésion partielle au programme ou les effets d'externalité peuvent aussi remettre en cause la validité interne des évaluations. Les dispositifs expérimentaux permettent, en général, de prendre en compte ces limites. Ils seront abordés dans la quatrième partie de l'article.

En termes de validité externe, les limites des évaluations proviennent de l'interprétation des résultats obtenus lorsque le programme ou la politique est testée dans un contexte particulier. En effet, ce qui est vrai dans un contexte peut ne pas l'être dans d'autres. Dans ce cas, la multiplication des évaluations dans des contextes différents (exemple du programme de rattrapage scolaire dans deux villes très différentes en Inde (cité dans la seconde section) permet de tester un même programme avec des conditions et des environnements différents (locations géographiques, équipe de mise en œuvre etc.) et analyser dans quelle mesure les résultats sont comparables (Banerjee et Duflo, 2008).

Par ailleurs, il peut être difficile d'extrapoler les résultats sur l'impact d'un programme évalué à petite échelle lorsque celui-ci est généralisé à un ensemble plus large. En effet, les évaluations ne prennent pas souvent en compte les effets d'équilibre général comme par exemple une saturation du marché dans le cas des programmes d'accompagnement vers l'emploi de différentes catégories de chômeurs[9]. Dans le cas de certaines

[9] *Dans un modèle simple de recherche d'emploi, Cahuc et Le Barbanchon (2008) analysent les conséquences de l'accompagnement fourni à des demandeurs d'emploi. Les auteurs observent que le fait de négliger les effets d'équilibre dus à ce programme peut mener à des conclusions erronées. En particulier, contrairement à ce que l'on pourrait penser, l'accompagnement peut augmenter le chômage à l'état stationnaire bien que les chercheurs d'emplois ont un plus fort taux de sortie du chômage que les non suivis. Ce type d'erreur est*

évaluations, on peut mettre en place des dispositifs spécifiques pour prendre en compte ces effets d'équilibre général. Par exemple, Crépon *et al.*[10], dans une étude en cours, évaluent l'impact d'un programme d'accompagnement des chômeurs diplômés dans un nombre important d'agences de l'emploi en France et randomisent l'intensité du traitement dans les différentes agences pour analyser les effets de saturation.

Enfin, il est important de considérer un ensemble d'hypothèses préalables sur l'impact du programme en dehors du contexte dans lequel il est évalué pour ne pas généraliser des résultats qui seraient liés à une situation particulière.

Concernant les limites spécifiques aux évaluations aléatoires, les conditions du dispositif expérimental peuvent limiter l'échantillon disponible pour l'étude. Les contraintes du dispositif expérimental peuvent amener à conduire l'évaluation sur une population spécifique, qui représente parfois seulement une certaine frange de la population cible, en choisissant par exemple les populations traitement et contrôle qui gênent le moins l'opérationnel, en n'incluant pas les individus les plus motivés, ou bien du fait que la méthode de randomisation s'applique seulement à une population particulière. Par exemple, Karlan et Zinman (2006) évaluent l'impact du crédit à la consommation en Afrique du Sud. Ils randomisent, dans « la bulle », certains individus qui ont un score de crédit juste en dessous du score minimal pour obtenir un crédit. Ces personnes sont assignées aléatoirement au traitement ou au contrôle. De ce fait, seul un effet local du programme est évalué et cela constitue un problème évident de validité externe. Cependant, cette limite s'applique systématiquement aux méthodes d'évaluation par régression en discontinuité (qui évaluent l'impact seulement sur ceux qui se situent juste au dessus ou en dessus du seuil).

Enfin, les expérimentations aléatoires ne peuvent être mises en œuvre que s'il existe une très forte collaboration entre les opérationnels et les responsables de l'évaluation et ce, dès le départ. Les évaluations aléatoires ne peuvent donc pas être effectuées *ex-post*. De ce fait, tous les projets ne peuvent donc pas être évalués.

*particulièrement présent quand la taille du groupe de traitement est faible.*

*[10] Crépon B, E.Duflo, M Gurgand, R. Rathelot et P. Zamora. Evaluation d'un programme d'accompagnement des jeunes chômeurs diplômés en France. Evaluation en cours menée pour le compte de l'ANPE.*

# Les évaluations aléatoires en pratique

Si ces évaluations reposent toutes sur l'assignation aléatoire du programme (avec un groupe traitement et un groupe contrôle sélectionnés aléatoirement), il existe en fait une diversité de dispositifs expérimentaux permettant d'évaluer différents types de programmes avec des modes d'interventions différents. D'autre part, même si le principe des évaluations aléatoires est très simple, mettre en place de tels dispositifs n'est pas forcément facile.

Deux types de faisabilité sont distingués dans cette section : une faisabilité *opérationnelle*, *i.e.* la possibilité de mettre en œuvre un protocole expérimental pour une politique ou une intervention donnée (en collaboration avec les opérationnels) et une faisabilité *statistique*, c'est-à-dire la capacité du dispositif expérimental à détecter les effets du programme.

## Conditions opérationnelles et stratégies de randomisation

La capacité à s'insérer dans le processus opérationnel d'une intervention est une condition nécessaire à la mise en œuvre des évaluations. La collaboration entre les équipes de recherche et les équipes opérationnelles responsables de l'intervention est primordiale à l'acceptation, à l'application et à la stabilité dans le temps du dispositif expérimental.

Ensuite, les dispositifs expérimentaux peuvent s'insérer avec plus ou moins de souplesse dans le fonctionnement des programmes. Dans la plupart des cas, les évaluations portent sur le pilote d'une intervention (qui sera généralisé en cas de résultats probants). La sélection aléatoire au sein des bénéficiaires ou groupes de bénéficiaires éligibles du pilote devient alors assez naturelle (cas par exemple de Progresa au Méxique). De la même façon, un programme qui a pour objectif de toucher l'ensemble de la population s'étend en général de manière progressive et constitue une opportunité intéressante pour les évaluateurs. En effet, la randomisation peut se faire au niveau de l'ordre de traitement des bénéficiaires même si tout le monde profitera du programme à la fin. En général les programmes d'interventions scolaires ne peuvent pas toucher toutes les écoles en même temps. Il est donc possible de profiter de l'extension progressive des programmes pour l'évaluer (déparasitage au Kenya). En règle générale (en dehors des évaluations aléatoires), le choix des bénéficiaires ou de l'ordre des bénéficiaires, est souvent basé sur des critères plus subjectifs qui peuvent être moins

justes (comme par exemple les liens politiques des communautés éligibles).

Il peut être impossible (ou difficile politiquement) de limiter l'accès des membres du groupe contrôle au programme. Le programme peut alors être ouvert à tout le monde, mais certaines personnes sont spécifiquement encouragées à participer (groupe de traitement) alors que d'autres personnes ne le sont pas. Dans ce cas, personne n'est exclu mais on s'attend à ce qu'il y ait une proportion plus importante de personnes traitées dans le groupe qui a été encouragé. La question de la validité statistique de ce type de dispositif est abordée par la suite.

Dans certains cas, il est difficile de randomiser, par exemple lorsque le nombre de personnes pouvant être servies est inférieur ou égal au nombre de personnes éligibles. De plus, les évaluations aléatoires peuvent poser des problèmes en terme d'éthique lorsqu'il existe une conviction forte que le programme a des effets positifs et qu'on refuse l'accès de certaines personnes qui auraient pu avoir accès, en absence d'évaluation, ou inversement que le programme a des effets négatifs connus et qu'il est quand même testé. Mais ces conditions sont rarement remplies. D'une part, une nouvelle intervention peut rarement toucher toute la population éligible au démarrage et dans de nombreux cas, les impacts du programme ne sont pas connus à l'avance. Enfin, la plupart des dispositifs expérimentaux des évaluations doivent être validés par des comités d'éthique au sein des universités auxquelles appartiennent les chercheurs impliqués dans l'évaluation.

Certains programmes ne peuvent cependant pas être évalués avec la méthode aléatoire lorsqu'on ne peut pas sélectionner à un niveau suffisamment désagrégé (si on ne peut randomiser qu'au niveau d'un faible nombre de régions) ou dans certaines conditions où randomiser poserait de vrais problèmes éthiques (comme par exemple l'aide alimentaire dans un contexte de famine).

## Niveau de randomisation

La sélection aléatoire des groupes traitement et contrôle peut se faire à plusieurs niveaux. La sélection au niveau individuel est la plus efficace en terme de capacité du dispositif de détecter les effets du traitement à détecter des effets d'un traitement, même s'il s'avère que les effets sont petits (cf. point suivant) mais peut s'avérer compliquée à mettre en place en pratique ou inappropriée étant donnée l'intervention évaluée. La sélection peut également se faire au niveau de groupes d'individus (villages, écoles, communautés).

Le niveau de randomisation dépend en général de l'intervention. Dans certains cas, l'effet d'un programme est mesuré sur un groupe plutôt que sur un individu. Par exemple, dans le cas d'évaluations au niveau des écoles (le degré de fréquentation des écoles, l'assiduité scolaire) ou de villages (existence de biens publics ou d'infrastructures), etc. Mais dans un grand nombre d'interventions, il y a une certaine flexibilité dans le niveau de randomisation. Le choix du niveau de randomisation a des conséquences importantes sur le budget et le dispositif de l'évaluation. Il est en effet plus efficace de randomiser au niveau individuel quand cela est faisable. Le niveau de randomisation est aussi à prendre en compte dans l'analyse des effets d'externalité lorsqu'il y a des biais liés aux comparaisons entre traitement et contrôle.

Par exemple, Miguel et Kremer (2004) ont évalué l'impact du déparasitage sur les élèves en randomisant au niveau des classes et ont détecté des effets beaucoup plus importants que les autres évaluations qui étaient randomisées au niveau individuel. Cette différence s'explique en partie par l'existence d'externalités possibles du traitement (les contrôles bénéficiant du fait que les traitements sont moins infectés). Même si ces externalités ne disparaissent pas totalement lorsque la randomisation est effectuée au niveau des écoles, elles sont beaucoup plus faibles.

Une autre forme d'externalité intervient lorsque les individus du groupe de comparaison changent leur comportement s'ils savent qu'ils vont être traités dans le futur. Là aussi, la randomisation au niveau d'un groupe (village ou école) permet de limiter la diffusion d'information.

## Validité statistique

La validité du dispositif expérimental, c'est-à-dire sa capacité à détecter les effets du programme lorsque ceux-ci existent, réside dans *sa puissance statistique*. La puissance d'une évaluation ou la capacité du dispositif à détecter des effets dépend d'un ensemble de paramètres[11] :

- la proportion de personnes participant parmi la population effectivement assignée au programme (la puissance augmentant avec le taux de participation) ;
- le niveau de l'effet que l'évaluation cherche à détecter. Plus le dispositif cherche à détecter des effets de petites tailles et plus la puissance de l'expérimentation est faible. La puissance de l'expérimentation doit donc être élevée pour

[11] *Pour une exposition formalisée de la puissance statistique des évaluations, voir Duflo et al. (2007).*

pouvoir détecter des effets les plus faibles du programme[12] ;

- Le nombre d'unités (individus ou groupes d'individus) randomisées. La puissance augmente avec le nombre d'unités qui sont randomisées. Lorsque la randomisation est effectuée au niveau de groupes d'individus (villages, communautés, classes etc.), la puissance dépend aussi du nombre d'individus suivis au sein des groupes et est une fonction croissante de ce nombre. Elle a, toutefois, une limite inférieure à un lorsque le nombre d'individus devient très grand. Ceci est lié à l'existence d'une composante spécifique au groupe. La puissance décroît avec l'importance relative de la composante spécifique au groupe.
- Enfin, la puissance de l'expérimentation décroît avec la variance des variables d'output d'intérêt.

Le choix du dispositif d'une évaluation doit prendre en compte l'ensemble de ces paramètres pour mesurer sa capacité à détecter les effets du programme. Certains paramètres peuvent être connus *ex-ante* (comme le nombre d'unités à randomiser ou les effets que l'évaluation cherche à détecter) alors que d'autres sont plus difficiles à estimer à l'avance. Ce sont le cas par exemple de la variance des variables d'output ou de la composante spécifique au groupe en l'absence de données disponibles. D'autre part, les dispositifs des évaluations sont souvent contraints en terme de budget ou d'unités disponibles pouvant être randomisées, d'autant plus lorsqu'il s'agit de groupes (villages ou communautés).

Il existe également des arbitrages à effectuer entre puissance statistique et faisabilité opérationnelle. Il est beaucoup plus simple d'atteindre un niveau de puissance statistique satisfaisant en randomisant au niveau de l'individu, mais il peut s'avérer politiquement ou pratiquement compliqué à mettre en place. En effet, il est difficile de randomiser l'assignation du traitement au sein d'une même communauté ou village (accès au micro-crédit par exemple). A l'inverse, les méthodes de randomisation par encouragement sont politiquement plus faciles à « vendre » mais comportent le risque d'une faible différence de taux de participation entre groupes traitement et contrôle (si l'encouragement ne fonctionne pas de manière satisfaisante et si une proportion importante de contrôle participe au programme).

Ensuite, beaucoup d'évaluations font face à des problèmes d'adhésion partielle au programme : certaines personnes du traitement ne reçoivent pas le programme et/ou certaines personnes du contrôle le reçoivent. Si certains programmes comme des interventions au niveau d'écoles ou la construction d'infrastructures traitent l'ensemble de la population effectivement assignée au traitement, la participation à de nombreux programmes est volontaire. Seulement une fraction des individus effectivement assignés au traitement participera au programme. Banerjee *et al.,* (2008) montrent que sur un projet en Inde ciblant les plus pauvres, une proportion significative de ménages (18 %) a refusé l'actif productif (vaches, chèvres etc.) pourtant offert sans contrepartie. Dans les programmes de micro-crédit, il est assez fréquent que seule une minorité de la population cible prenne un crédit. L'adhésion partielle peut réduire significativement la puissance puisque certains individus du groupe traitement n'auront finalement pas bénéficié du programme. De ce fait la taille des effets mesurés est différente selon que l'adhésion partielle au programme est prise en compte ou non (voir la section suivante sur la présentation des résultats). Il faut donc bien prendre en compte le taux d'adhésion au programme dans le choix de la taille de l'échantillon.

## Stratification

Enfin, il est possible que, malgré la randomisation, les groupes traitement et contrôle n'aient pas les mêmes caractéristiques et ceci uniquement dans le cas de petits échantillons. Bruhn et McKenzie (2008) décrivent et analysent les différentes méthodes utilisées pour minimiser ces différences sur des variables observables. Les deux principales méthodes sont la stratification et l'appariement par paire. Ces méthodes permettent de constituer des groupes traitement et contrôle plus équilibrés en terme de variables observables lorsque la randomisation s'effectue sur un nombre limité d'unités.

La stratification correspond à la division de l'échantillon en plusieurs sous échantillons différents selon des variables observables. A l'intérieur de chaque sous échantillon, il s'agit de sélectionner un nombre égal d'observations traitement et contrôle. La randomisation a donc lieu au sein de chaque sous échantillon. L'objectif de la stratification est d'équilibrer l'échantillon par rapport à certaines variables clés. La stratification est également importante si les effets du programme sont susceptibles d'être différents sur des sous-groupes d'échantillon, il faut cependant avoir suffisamment d'individus dans chaque sous-groupe pour mesurer l'hétérogénéité du traitement. Les variables utilisées pour la stratification sont en général fortement corrélées avec la variable

*[12] En général, les effets considérés comme étant faibles se situent aux alentours de 0,2 écarts types, les effets moyens à environ 0,4 écarts types et des effets élevés aux environs de 0,6 écarts types.*

d'output ou devraient avoir en théorie un impact important sur le résultat.

Cependant, il peut être complexe de stratifier sur plusieurs variables. Non seulement le tirage aléatoire peut apparaître moins transparent (surtout lorsqu'il est effectué sur place, par exemple lors de loteries en public). Il diminue aussi le nombre de degrés de liberté car il faut ensuite intégrer les variables de stratification dans les régressions d'impact.

Une version très avancée de stratification est l'appariement par paire où des paires d'unités (individus, villages etc.) sont constituées en fonction d'un nombre de caractéristiques observables ressemblantes et, au sein de chaque paire, une unité est aléatoirement assignée au groupe traitement et l'autre au groupe contrôle. Cette méthode permet d'équilibrer les groupes pour un plus grand nombre de variables observables et, constitue une des meilleures méthodes pour obtenir l'équilibre dans des échantillons de faibles tailles. (Bruhn et McKenzie, 2008).

## Présentation des résultats

La présentation des résultats des évaluations d'impact est simple et a le mérite d'être très transparente. La mesure la plus simple d'une évaluation est le calcul de la différence sur les variables d'outputs entre les groupes traitement et contrôle.

$$ITT = \bar{y}(1) - \bar{y}(0) \qquad (3)$$

L'intention de traiter (ITT) mesure l'effet global du traitement. En effet, même si une proportion des bénéficiaires n'a pas accès au programme, l'intention de traiter mesure l'effet de la politique dans son ensemble sur la population cible (en prenant en compte le fait que certains, assignés au traitement, participent ou ne participent pas). Elle mesure globalement l'effet de la politique ou du programme sur la population cible. Cette mesure fournit une estimation de l'impact réel de l'offre du programme. Dans beaucoup de programmes, qu'il s'agisse de campagnes de vaccination, de transferts d'actifs productifs pour les plus pauvres ou de politiques d'accompagnement des chercheurs d'emploi, il est rare que l'ensemble de la population cible participe. Il est cependant intéressant d'estimer l'effet global de la politique en prenant en compte que seule une proportion de la population cible participe, car ce sera le cas dans la réalité si le programme est généralisé.

L'impact peut aussi être mesuré sur ceux qui ont effectivement participé au programme.

Il s'agit alors de calculer la différence sur les variables d'outputs entre les groupes traitement et contrôle et de la diviser par la proportion d'individus ayant effectivement reçu le programme parmi ceux auxquels le programme a été assigné et parmi ceux auxquels il n'a pas été assigné. Il s'agit là de mesurer l'effet du traitement sur les traités (TT) :

$$TT = \frac{\bar{y}(1) - \bar{y}(0)}{P(C=1|T=1) - P(C=1|T=0)} \qquad (4)$$

La simplicité de cet estimateur illustre la valeur du cadre expérimental. La population des unités de contrôle se décompose comme dans les unités traitement en une population ayant accès au programme et une population n'ayant pas accès. La seule différence est que comme le programme n'y est pas proposé, ces deux populations ne sont pas identifiées, mais le point important est qu'elles sont semblables à celles des unités traitement. L'intérêt de l'estimation TT est de mesurer l'impact du programme sur ceux qui en ont effectivement bénéficié. Il s'agit par exemple de l'effet du micro-crédit sur ceux qui en ont effectivement reçu (sachant que la participation est volontaire).

Les deux mesures, ITT et TT, sont intéressantes selon que l'on souhaite mesurer l'effet du programme dans son ensemble ou seulement sur ceux qui en ont bénéficié.

## Conclusion

Les évaluations aléatoires ont connu un engouement important au cours des dix dernières années, elles représentent une avancée réelle dans l'analyse de l'impact des programmes sociaux ou de développement. Alors que ces évaluations ont démarré aux Etats-Unis et en Europe du nord, leur véritable expansion s'est effectuée dans les pays en développement. Même si elles représentent encore une faible proportion des évaluations d'impact, elles sont appliquées à un champ d'interventions de plus en plus large. L'accumulation des résultats des évaluations permet de tirer des conclusions sur l'impact de certaines interventions : sur des programmes de santé, d'éducation et bientôt dans d'autres secteurs où il existe peu d'évaluations rigoureuses comme la microfinance par exemple. Les évaluations aléatoires ne se limitent pas seulement à l'impact d'un programme, elles permettent aussi de comparer plusieurs modalités d'une intervention (et voir laquelle est la plus efficace), de tester des innovations du processus des

interventions ou encore d'analyser des hypothèses de la théorie économique.

La multiplication des évaluations montre qu'un grand nombre d'interventions sont éligibles et qu'il est possible de travailler avec de nombreux partenaires au niveau des ONG, des gouvernements et des organisations internationales. D'autre part, l'application de la méthode d'évaluation aléatoire à de nombreux contextes et la diversité des dispositifs expérimentaux ont permis aux chercheurs de prendre en compte certaines de ses limites.

Les évaluations aléatoires ont un rôle important à jouer pour informer et guider les décideurs politiques et les responsables de programmes sur la base de preuves empiriques rigoureusement construites.

## Références Bibliographiques


**Banerjee A, Duflo E. (2008),** « The Experimental Approach to Development Economics ». Massachusetts Institute of technology, Department of Economics and Abdul Latif Jameel Poverty Action Lab.

**Banerjee A, Cole S., Duflo E., Linden L. (2007),** « Remedying Education : Evidence from Two Randomized Experiments in India », *Quarterly Journal of Economics*, Vol. 122-3, pp 1235-1264.

**Banerjee A, Duflo E., Chattopadhyay R., Shapiro J., (2007),** « Targeting efficiency : How Well Can We Identify the Poor », Institute for Financial Management and Research. Centre for Microfinance. Working Paper Series n°21.

**Bruhn M., McKenzie D. (2008),** « In pursuit of balance : randomization in practice in development field experiments ». World Bank Policy Research Working Paper Series n°4752.

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

**Buddelmeyer H, Skoufias E. (2003),** « An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA ». IZA Working Paper n°827.

**Cahuc P., Le Barbanchon T. (2008),** « Labor Market Policy Evaluation in Equilibrium : Some Lessons of the Job Search and Matching Model ». IZA Discussion Paper Series n°3687. September 2008.

**Cohen J., Dupas P. (2007),** « Free Distribution or Cost-Sharing? Evidence from a randomized malaria prevention experiment, » Brookings Institution Global Working Paper n°14.

**Crépon B, Devoto F., Duflo E., Parienté W. (2008),** « Poverty, Access to Credit and the Determinants of Participation to a New Micro-credit Program in Rural Areas of Morocco », Ex Post Impact Analyses Series n 2, Agence Française de Développement, Paris.

**Duflo E, Dupas P., Kremer M., Sinei S. (2006),** « Education and HIV/AIDS Prevention : Evidence from a randomized evaluation in Western Kenya », World Bank Policy Research Working Paper n°402.

**Duflo E., Kremer M., Glennerster R. (2007),** « Using Randomization in Development Economics Research : A Toolkit, » in *Handbook of Development Economics*. Elsevier-North Holland John Strauss and Paul Schultz, editors, Vol. 4.

**Gertler P. J., Boyce S. (2001),** « An Experiment in Incentive-Based Welfare : The Impact of PROGRESA on Health in Mexico » Mimeo, UC-Berkeley.

**Giné X, Karlan D. (2006),** « Group versus individual liability : evidence from a field experiment in the Philippines », Yale University Economic Growth Center working paper n°940.

**Glazerman S., Levy D., Myers D. (2003),** *Non experimental Replications of Social Experiments : A systematic Review.* Princeton, NJ : Mathematical Policy Research, Inc.

**Heckman J., Ichimura H., Todd P. (1997),** « Matching as an Econometric Evaluation Estimator : Evidence from Evaluating a Job Training Program », *Review of Economic Studies*, October 1997, Vol°64-4, pp 605-654.

**Imbens W., Wooldridge J. (2008),** « Recent Developments in the Econometrics of Program Evaluation ». NBER Working Paper n°W14251.

**International Food Policy Research (2000),** « Monitoring and Evaluation System », Discussion paper, International Food Policy Research (IFPRI).

**J-PAL (2007),** « Mass Deworming : A Best-Buy for Education and Health », Policy Briefcase n°4. http://www.povertyactionlab.org/papers/Briefcase%204%20deworming.pdf.

**Karlan D., Zinman J. (2005),** « Observing Unobservables : Identifying Information Asymmetries with a Consumer Credit Field Experiment », BREAD Working Paper n° 94.

**Karlan D., Zinman J. (2007),** « Expanding Credit Access : Using Randomized Supply Decisions to Estimate the Impacts », MIMEO, Yale University.

**Kremer M., Miguel E. (2004),** « Worms : Identifying Impacts on Education and Health in the Presence of Treatment Externalities », *Econometrica*, Vol. 72-1, pp 159-217.

**Kremer, M., Miguel E. (2007),** « The Illusion of Sustainability », *Quarterly Journal of Economics* Vol. 112-3, pp 1007-1065.

**LaLonde R.J. (1986),** « Evaluation the Economic Evaluations of Training Programs Using Experimental Data », *American Economic Review*, Vol. 76-4, pp 602-620.

**Maluccio J. A., Flores R. (2005),** « Impact Evaluation of a Conditional Cash Transfer Program », Discussion paper, International Food Policy Research Institute, Research Report n°141.

**Morduch J. (1998),** « Does Microfinance Really Help the Poor ? New Evidence from Flagship Programs in Bangladesh », Mimeo, Princeton University.

**Olken B. (2006),** « Corruption and the Costs of Redistribution : Micro Evidence from Indonesia », *Journal of Public Economics*, Vol. 90-4/5, pp 853-870.

**Olken B. (2007),** « Monitoring Corruption : Evidence from a Field Experiment in Indonesia », *Journal of Political Economy*, Vol 115-2, pp 200-249.

**Pitt M., Khandker S (1998),** « The Impact of Group-Based Credit Programs on Poor Households in Bangladesh : Does the Gender of Participants Matter ? » *The Journal of Political Economy*, Vol. 106-5, pp 985-996.

**Rubin D.B. (1974),** « Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies », *Journal of Educational Psychology*, Vol. 66, pp 688-701.

**Schultz T. P. (2004),** « School subsidies for the poor : Evaluating the Mexican PROGRESA poverty program », *Journal of Development Economics*, Vol. 74-1, pp 199-250.

**WHO (2004),** « Action Against Worms » Partners for Parasite Control Newsletter, Issue 1, January 2004 (available at www.who.int/wormcontrol/en/action_against_worms.pdf)