Évaluer l'impact des programmes d'aide au développement : le rôle des évaluations par assignation aléatoire

Evaluating the Impact of Development Aid Program : The Role of Randomized Evaluations

Esther Duflo*

Massachusetts Institute of Technology, Department of Economics and Poverty Action Lab

Parce que les programmes dont on a montré qu'ils sont une réussite peuvent être reproduits dans d'autres pays, alors que les programmes ayant échoué peuvent être abandonnés, les évaluations d'impact sont des biens publics internationaux : les agences internationales devraient donc jouer un rôle clé dans leur promotion et leur financement. En agissant ainsi, elles pourraient atteindre trois objectifs importants : améliorer les taux de rendement des programmes qu'elles supportent ; améliorer les taux de rendement des programmes que les autres décideurs politiques supportent, en fournissant les observations sur la base desquelles les programmes peuvent être sélectionnés ; construire un support de long terme pour l'aide internationale et le développement, en rendant possible le fait de signaler de manière crédible les programmes qui fonctionnent et ceux qui ne fonctionnent pas.

L'article avance qu'il existe des possibilités considérables pour étendre l'utilisation des évaluations par assignation aléatoire. Pour une large classe de programmes de développement (quoique pas pour tous), les évaluations par assignation aléatoire peuvent être utilisées pour surmonter les problèmes souvent rencontrés dans l'utilisation pratique des évaluations. Premièrement, il traite de la méthodologie de l'évaluation par assignation aléatoire à travers plusieurs exemples concrets, tirés pour

* Je remercie Abhijit Banerjee, Angus Deaton, Michael Kremer, Edward Miguel et Rachel Glennerster pour leurs commentaires utiles. L'article a bénéficié de commentaires de Paul Shultz and Francois Bourguignon sur un article lié présenté à la conférence ABCDE à Bengalore (« Le passage à grande échelle et l'évaluation ») et de travaux sur l'article « Utiliser l'assignation aléatoire pour l'évaluation de l'efficacité du développement », écrit en collaboration avec Michael Kremer et présenté à la conférence OED à Washington. Je suis très reconnaissante du support financier de la fondation Alfred P. Sloan et du *National Institute of Health*.

la plupart du cas de l'Inde. Il débat alors du potentiel de l'évaluation par assignation aléatoire en tant que base pour étendre les projets. Enfin, il traite des pratiques courantes et du rôle que les agences internationales peuvent jouer dans la promotion et le financement d'évaluations rigoureuses.

Classification JEL : F35, O19.

Because programs that have been shown to be successful can be replicated in other countries, while unsuccessful programs can be abandoned, impact evaluations are international public goods : the international agencies should thus have a key role in promoting and financing them. In doing this, they would achieve three important objectives : improve the rates of returns on the programs they support; improve the rates of returns on the programs other policymakers support, by providing evidence on the basis of which programs can be selected; build long term support for international aid and development, by making it possible to credibly signal what programs work and what programs do not work.

The paper argues there is considerable scope to expand the use of randomized evaluations. For a broad class of development programs (although not all of them), randomized evaluation can be used to overcome the problems often encountered when using evaluation practices. First, it discusses the methodology of randomized evaluation through several concrete examples, mostly drawn from India. It then discusses the potential of randomized evaluation as a basis for scaling up. Finally, it discusses current practices and the role international agencies can play in promoting and financing rigorous evaluations.

INTRODUCTION

Les bénéfices de l'évaluation de l'impact des programmes de développement et de la connaissance des programmes qui fonctionnent et ceux qui ne fonctionnent pas ne dépassent pas beaucoup un programme ou une agence quelconque. Les évaluations d'impact crédibles sont des biens publics mondiaux dans le sens où elles peuvent être des guides fiables pour les organisations internationales, les gouvernements, les donneurs et les ONG au-delà des frontières nationales. Le manque d'évaluation crédible des programmes de développement est aussi souvent lu par le public comme un signe de la défaillance des programmes eux-mêmes, à tel point qu'une évaluation d'impact crédible peut aider à construire un support pour le développement.

Les méthodes traditionnelles de mesure de l'impact d'un programme peuvent être sujettes à des biais sérieux dus à des variables omises. Pour une large catégorie de programmes de développement, les évaluations par assignation aléatoire peuvent être utilisées pour aborder ces problèmes. Tous les programmes ne peuvent bien évidemment pas être évalués par les évaluations par assignation aléatoire; par exemple, l'examen de problématiques telles que celle de l'indépendance de la

banque centrale doit reposer sur d'autres méthodes d'évaluation. Les programmes visant les individus ou les communautés locales (tels que l'hygiène, les réformes du gouvernement local, l'éducation et la santé) ont des chances d'être candidats aux évaluations par assignation aléatoire; cet article utilise le cas des programmes d'éducation dans les pays en développement comme exemple.

Historiquement, les évaluations prospectives par assignation aléatoire de programmes de développement ont constitué une minuscule fraction de toutes les évaluations du développement. Dans cet article, nous soutenons qu'il y a de la marge pour étendre leur utilisation, bien qu'elles doivent nécessairement rester une petite fraction de toutes les évaluations.

Nous ne proposons pas que tous les projets soient sujets aux évaluations par assignation aléatoire. Mais nous soutenons qu'il y a actuellement un énorme déséquilibre dans la méthodologie d'évaluation, et qu'accroître la part des projets sujets à l'évaluation aléatoire de presque rien à une fraction même petite pourrait avoir un formidable impact sur la connaissance de ce qui fonctionne dans le développement. La politique de développement est beaucoup trop souvent fondée sur des lubies et les évaluations par assignation aléatoire pourraient lui permettre d'être fondée sur des preuves.

L'article procède comme suit : la section 1 discute de la méthodologie des évaluations par assignation aléatoire : nous présentons le problème de l'évaluation d'impact, nous passons en revue les raisons pour lesquelles les autres méthodes d'évaluation actuelles peuvent souvent être incapables de contrôler de manière adéquate pour le biais de sélection et nous exposons les raisons pour lesquelles les évaluations par assignation aléatoire peuvent être utiles pour aborder les problèmes rencontrés par les autres pratiques d'évaluation. La section 2 passe en revue les évaluations par assignation aléatoire récentes dans les pays en développement. La section 3 tire les leçons des évaluations décrites dans la section 2 et la section 4 passe en revue un exemple de pratique courante, offre des explications de politique économique au fait que les évaluations par assignation aléatoire soient si rares et discute du rôle que les agences internationales peuvent jouer dans la promotion et le financement d'évaluations rigoureuses, y compris les évaluations par assignation aléatoire. La section 5 conclut.

1 LA MÉTHODOLOGIE DE L'ÉVALUATION ALÉATOIRE

Le problème de l'évaluation

Toute évaluation d'impact tente essentiellement de répondre à une question de contrefactuel : comment un individu qui n'a pas bénéficié du programme aurait-il été en l'absence du programme ? Comment ceux qui n'ont pas bénéficié du programme auraient-ils été s'ils avaient été exposés au programme ? La difficulté avec ces questions est immédiate : à un moment donné, un individu est observé soit exposé au programme, soit non exposé. Comparer le même individu dans le temps ne nous donnera pas, dans la plupart des cas, une estimation fiable de l'impact que le programme a eu sur lui, puisque beaucoup d'autres choses ont pu changer en même temps que le programme a été introduit. Nous ne pouvons donc pas chercher à obtenir une estimation de l'impact du programme sur chaque individu. Nous pouvons uniquement espérer être capables d'obtenir l'impact moyen du programme sur un groupe d'individus en les comparant à un groupe similaire qui n'a pas été exposé au programme. L'objectif critique d'une évaluation d'impact est donc d'établir un *groupe de comparaison* crédible, un groupe d'individus qui, *en l'absence du programme,* auraient eu des résultats similaires à ceux qui ont été exposés au programme. Ce groupe nous donne une idée de ce qui se serait passé pour le groupe du programme s'il n'avait pas été exposé et nous permet donc d'obtenir une estimation de l'impact moyen sur le groupe en question. En général, dans le monde réel, les individus qui ont été sujets au programme et ceux qui ne l'ont pas été sont très différents : les programmes sont placés dans des zones spécifiques (par exemple, des zones plus ou moins riches), les individus sont filtrés pour participer au programme (par exemple, sur la base de la pauvreté ou sur la base de leur motivation) et, enfin, la décision de participer est souvent volontaire. Pour toutes ces raisons, ceux qui n'ont pas été exposés au programme sont souvent un mauvais groupe de comparaison pour ceux qui l'ont été. Toute différence entre eux peut être attribuée à deux facteurs : des différences pré-existantes (ledit « biais de sélection ») et l'impact du programme. Puisque nous n'avons aucun moyen fiable d'estimer la taille du biais de sélection, nous ne pouvons pas décomposer la différence totale entre un effet de traitement et un terme de biais.

Pour résoudre ce problème, les évaluations de programme ont spécifiquement besoin d'être soigneusement planifiées à l'avance afin de déterminer le groupe qui puisse être un groupe de contrôle. Une situa-

tion dans laquelle le biais de sélection disparaît est celle dans laquelle les groupes de traitement et de comparaison sont sélectionnés de manière aléatoire au sein d'une population de bénéficiaires potentiels (des individus, des communautés, des écoles ou des salles de classe peuvent être sélectionnés dans le programme). Dans ce cas, en moyenne, nous pouvons être assurés que ceux qui sont exposés au programme ne sont pas différents de ceux qui ne le sont pas et qu'une différence statistiquement significative entre eux dans les résultats que le programme avait prévu d'atteindre après sa mise en place peut être attribuée au programme en toute confiance.

Comme nous le verrons plus loin dans cet article, l'assignation aléatoire des groupes de traitement et de comparaison peut intervenir en plusieurs circonstances. En prenant l'exemple de PROGRESA, un programme destiné à accroître la participation scolaire au Mexique, nous exposons comment les évaluations par assignation aléatoire prospectives peuvent être utilisées et comment leurs résultats peuvent aider à échelonner les programmes avec succès; en prenant les exemples de programmes de santé basés dans les écoles au Kenya et en Inde, nous illustrons comment les évaluations par assignation aléatoire prospectives peuvent être utilisées lors de la mise en œuvre de répliques adaptées des programmes; et, en prenant l'exemple des quotas pour les femmes politiques en Inde, nous discutons des façons d'utiliser l'assignation aléatoire induite par les programmes.

Cela vaut la peine d'apporter quelques clarifications concernant l'utilisation d'évaluations par assignation aléatoire pour estimer les effets d'un programme. Premièrement, une distinction peut être faite concernant ce que l'évaluation tente précisément d'estimer. Les évaluations par assignation aléatoire peuvent être utilisées pour estimer l'effet d'un traitement soit sur l'ensemble de la population qui a été sujette à l'assignation aléatoire, soit sur un sous-ensemble de la population défini par des caractéristiques prédéterminées, alors que les techniques de variables instrumentales estiment les effets de traitement sur les moyennes locales. Deuxièmement, les évaluations par assignation aléatoire estiment les effets de traitement en équilibre partiel, qui peuvent être différents des effets de traitement en équilibre général. Il est possible que, si quelques programmes d'éducation sont mis en place sur une large échelle, les programmes puissent affecter le fonctionnement du système éducatif et donc avoir un impact différent.

Autres techniques pour contrôler pour la sélection et d'autres biais de variables omises

Les évaluations par assignation aléatoire naturelles ou organisées ne sont pas la seule méthodologie qui puisse être utilisée pour obtenir des évaluations d'impact crédibles des effets des programmes. Les chercheurs ont développé des techniques alternatives pour contrôler pour les biais aussi bien que possible et des progrès ont été accomplis, notamment par les économistes du travail[1]. Dans ce qui suit, nous passons brièvement en revue quelques-unes des techniques les plus populaires parmi les chercheurs : l'appariement par score de propension, les estimations par différences des différences et le modèle de discontinuité des régressions.

Une stratégie pour contrôler pour le biais est de tenter de trouver un groupe de contrôle qui soit aussi comparable que possible au groupe de traitement, au moins concernant les dimensions observables. Ceci peut être fait en collectant autant de variables explicatives que possible puis en ajustant les différences calculées à l'aide d'une régression ou en « appariant » les groupes du programme et de comparaison jusqu'à former un groupe de comparaison qui soit aussi similaire que possible au groupe du programme. Une possibilité est de prédire la probabilité qu'un individu donné soit dans le groupe de comparaison ou de traitement sur la base de toutes les caractéristiques observables disponibles, puis de former un groupe de comparaison en sélectionnant les personnes ayant la même probabilité d'être traitées que celles qui sont effectivement traitées (« appariement par score de propension »). Le défi de cette méthode et des contrôles de régression est qu'elle s'articule autour du fait d'avoir identifié toutes les différences potentiellement pertinentes entre les groupes de traitement et de contrôle. Dans les cas où le traitement est assigné sur la base d'une variable qui n'est pas observée par les chercheurs (la demande pour le service par exemple), cette technique peut mener à des déductions trompeuses.

Une seconde stratégie est ce que l'on appelle souvent la technique de la « différence dans la différence ». Quand on peut correctement argumenter le fait que le résultat n'aurait pas connu de tendances différentielles dans les régions qui ont reçu le programme si le programme n'avait pas été mis en place, il est possible de comparer la croissance des variables d'intérêt entre les régions du programme et celles hors programme. Cependant, il est important de ne pas prendre cette hypothèse comme étant donnée. Cette hypothèse d'identification ne peut

1 Il y a de nombreuses et excellentes revues techniques et non techniques de ces méthodes ainsi que de leur valeur et leurs limites. Voir Angrist et Krueger, 1999 et 2001; Card, 1999; et Meyer, 1995.

pas être testée, et même pour la vérifier de manière plausible, il faut disposer de séries temporelles longues pour les données antérieures à la mise en place du programme afin de pouvoir comparer les tendances sur des périodes suffisamment longues. Il faut également s'assurer qu'aucun autre programme n'a été mis en place en même temps (ce qui n'est pas souvent le cas). Enfin, il faut également prendre en compte, en tirant des déductions, le fait que les régions sont souvent affectées par des chocs persistants dans le temps qui peuvent ressembler à des « effets programme ». Bertrand, Duflo et Mullainathan (2002) ont trouvé que les estimations en différence dans les différences (telles qu'elles sont communément réalisées) peuvent durement biaiser les écarts-types : les chercheurs ont généré de manière aléatoire des lois placebo et ont trouvé qu'avec environ 20 ans de données, les estimations en différence dans les différences trouvent un « effet » significatif à 5 % pour pas moins de 45 % des lois placebo.

Pour illustrer les cas où les estimations en différence dans les différences peuvent être utilisées, Duflo (2001) a profité d'un programme d'expansion scolaire rapide mis en place en Indonésie dans les années 1970 pour estimer l'impact de la construction d'écoles sur la scolarisation et les salaires ultérieurs. L'identification a été rendue possible par le fait que la règle d'allocation pour les écoles était connue (plus d'écoles ont été construites dans les endroits avec de faibles taux de scolarisation initiaux) et par le fait que les cohortes participant au programme étaient facilement identifiables (les enfants âgés de 12 ans ou plus quand le programme a débuté n'y participaient pas). La croissance accrue de l'éducation entre cohortes dans les régions ayant reçu plus d'écoles suggère que l'accès aux écoles contribue à accroître l'éducation. Les tendances étaient plutôt parallèles avant le programme et se sont clairement inversées pour la première cohorte qui fut exposée au programme, renforçant par-là même la confiance dans l'hypothèse d'identification. Cependant, cette stratégie d'identification n'est habituellement pas valide; souvent, quand les changements de politique sont utilisés pour identifier les effets d'une politique particulière, le changement de politique est lui-même endogène au résultat qu'il était censé affecter, rendant du coup l'identification impossible (Besley et Case, 2000).

Enfin, une troisième stratégie , appelée le « modèle de discontinuité des régressions » (Campbell, 1969), profite du fait que les règles d'un programme engendrent parfois des discontinuités qui peuvent être utilisées pour identifier l'effet d'un programme en comparant ceux au dessus d'un certain seuil à ceux juste en dessous. Si les ressources sont allouées sur la base d'un certain nombre de points, il est possible de

comparer ceux juste au dessus à ceux juste en dessous du seuil. Angrist et Lavy (1999) utilisent cette technique pour évaluer l'impact de la taille des classes en Israël, où un second enseignant est alloué chaque fois que la taille d'une classe dépasse les 40. Cette politique engendre des discontinuités dans la taille des classes lorsque les inscriptions dans un niveau augmentent de 40 à 41 (puisque la taille de la classe passe d'une classe de 40 élèves à deux classes de 20 et 21 élèves chacune). Angrist et Lavy ont comparé les résultats des tests dans les classes juste au dessus et juste en dessous de ce seuil et ont trouvé que ceux juste au dessus du seuil obtenaient des résultats significativement plus élevés que ceux juste en dessous – ce qui peut être attribué en toute confiance à la taille des classes puisqu'il est très improbable que les écoles situées de chaque côté du seuil aient toute autre différence systématique[2]. De telles discontinuités dans les règles des programmes, quand elles sont appliquées, sont alors sources d'identification.

Dans les pays en développement, cependant, il est souvent possible que les règles ne soient pas appliquées de manière suffisamment stricte pour engendrer des discontinuités qui puissent être utilisées pour des motifs d'identification. Par exemple, les chercheurs ont tenté d'utiliser comme source d'identification la discontinuité introduite dans la politique de la banque Grameen (la plus importante organisation de micro crédit au Bangladesh) par le fait de ne prêter qu'aux personnes propriétaires de moins d'une acre de terrain (Land et Khandker, 1998). Il se trouve qu'en pratique, la banque Grameen prête à beaucoup de personnes possédant plus d'une acre de terrain et qu'il n'y a pas de discontinuité dans la probabilité d'emprunt au seuil (Morduch, 1998).

Ces trois techniques sont sujettes à de larges biais qui peuvent conduire soit à une surestimation soit à une sous-estimation de l'impact des programmes. LaLonde (1986) a trouvé que beaucoup des procédures économétriques et des groupes de comparaison utilisés dans les évaluations de programme ne donnent pas des estimations correctes ou précises et que des telles estimations économétriques diffèrent souvent de manière significative des résultats expérimentaux.

Les problèmes d'identification avec des méthodes d'évaluation sans assignation aléatoire doivent être abordés avec une extrême précaution parce qu'elles sont moins transparentes et plus sujettes à des divergences d'opinion que ces mêmes problèmes avec des évaluations par assignation

2 Angrist et Lavy notent que les parents qui découvrent qu'ils ont eu un mauvais tirage à la « loterie de l'inscription » (par exemple, 38 inscriptions) devraient alors retirer les enfants du système d'enseignement public pour les inscrire dans des écoles privées. Cependant, comme Angrist et Lavy en débattent, les cours élémentaires privés sont rares en Israël sortis des communautés ultra orthodoxes.

aléatoire. De plus, les différences entre les bonnes et les mauvaises évaluations sans assignation aléatoire sont difficiles à communiquer, particulièrement aux décideurs politiques, à cause de toutes les mises en garde qui doivent accompagner leurs résultats. En pratique, ces mises en garde ne devraient jamais être données aux décideurs politiques, et même si elles sont données, elles devraient être ignorées; dans les deux cas, les décideurs politiques ont des chances d'être induits en erreur. Cela suggère que, bien que les évaluations sans assignation aléatoire soient encore nécessaires, il devrait y avoir un engagement à mener des évaluations par assignation aléatoire lorsque c'est possible.

2 EXEMPLES D'ÉVALUATIONS PAR ASSIGNATION ALÉATOIRE PROSPECTIVES

Les projets pilotes

Avant qu'un programme ne soit lancé à grande échelle, un projet pilote, nécessairement d'envergure limitée, est souvent mis en place. Choisir de manière aléatoire les bénéficiaires du pilote peut être fait dans la plupart des circonstances puisque de nombreux sites potentiels (ou individus) méritent autant les uns que les autres d'être les lieux où le pilote prend place. Le pilote peut alors être utilisé, non seulement si le programme est faisable (ce qui est l'utilisation faite de la plupart des pilotes sur le moment), mais également quand le programme a les effets attendus. Les programmes d'apprentissage et de maintien du revenu étaient des exemples frappants des évaluations par assignation aléatoire. Un nombre croissant de tels projets pilotes sont évalués, souvent en collaboration entre une ONG et des universitaires (voir, par exemple, Kremer 2003 pour plusieurs références). Pour illustrer brièvement comment ces études peuvent fonctionner en pratique, nous analysons un exemple en Inde, analysé par Banerjee *et al.* (2001). Cette étude a évalué un programme dans lequel une ONG indienne (Seva Mandir) a décidé d'engager un second enseignant dans des centres d'éducation non formelle qu'elle gère dans des villages. Les écoles non formelles cherchent à fournir des connaissances basiques en lecture, écriture et calcul à des enfants qui ne vont pas à l'école formelle et, à moyen terme, à aider ces enfants à « revenir dans le droit chemin » du système scolaire régulier. Ces centres sont touchés par un très fort absentéisme des enseignants et des enfants. Un second enseignant (souvent une femme) a été assigné de manière aléatoire à 21 des 42 écoles. L'espoir était d'accroître le nombre de jours

d'ouverture de l'école, accroître la participation des enfants et accroître les performances en fournissant une attention plus individualisée aux enfants. En choisissant une enseignante, l'ONG espérait également rendre l'école plus attractive pour les filles. L'assiduité des enseignants et des enfants était régulièrement surveillée dans les écoles du programme et de comparaison pendant toute la durée du projet. L'impact du programme sur l'apprentissage était mesuré en testant les enfants à la fin de l'année scolaire. Le programme a réduit les nombre de jours de fermeture des écoles : les écoles avec un seul enseignant sont fermées 39 % du temps alors que les écoles avec deux enseignants sont fermées 24 % du temps. L'assiduité des filles a augmenté de 50 %. Cependant, il n'a eu aucune différence dans les résultats des tests.

Les projets pilotes évalués avec précaution constituent une base solide pour la décision de faire passer le projet au niveau supérieur. Dans l'exemple qui vient d'être discuté, le programme à deux enseignants ne fut pas mis en place à grande échelle par l'ONG, au motif que les bénéfices n'étaient pas suffisants pour dépasser les coûts. Les économies ont été utilisées pour étendre d'autres programmes. Des résultats positifs, d'un autre côté, peuvent aider à bâtir un consensus pour le projet, qui a le potentiel pour être étendu bien au-delà de l'échelle prévue initialement. Le programme PROGRESA au Mexique est l'exemple le plus frappant de ce phénomène. PROGRESA offre des bourses d'études, distribuées aux femmes, conditionnelles à l'assiduité des enfants et à des mesures de santé préventives (supplément alimentaire, visites médicales et participation à des programmes d'éducation sur la santé). En 1998, quand le programme a été lancé, des officiels du gouvernement mexicain ont consciemment pris la décision de profiter du fait que les contraintes budgétaires rendaient impossible le fait d'atteindre les 50000 communautés potentiellement bénéficiaires de PROGRESA en même temps et ont préféré commencer avec un programme pilote dans 506 communautés. La moitié d'entre elles ont été sélectionnées de manière aléatoire pour recevoir le programme et les données de base et subséquentes furent collectées dans les communautés restantes (Gertler et Boyce 2001). Une partie de la justification pour débuter avec ce programme pilote était d'accroître la probabilité que le programme continue en cas de changement du parti au pouvoir. Les partisans du programme comprirent que, pour être étendu avec succès, le programme aurait besoin d'un support politique continu. La tâche d'évaluer le programme fut donnée à des chercheurs universitaires, à travers l'Institut International de Recherche pour la Politique Alimentaire. Les données furent rendues accessibles à de nombreuses personnes différentes et de nombreux articles ont été écrits sur son impact (la plupart d'entre

eux sont disponibles sur le site Internet de l'IFPRI). Les évaluations ont montré que le programme était efficace pour améliorer la santé et l'éducation : en comparant les bénéficiaires et les non bénéficiaires de PROGRESA, Gertler et Boyce (2001) montrent que les enfants connaissent environ une réduction de 23 % dans l'incidence de la maladie, une augmentation de 1-4 % de leur taille et une réduction de 18 % de l'anémie. Les adultes connaissent une réduction de 19 % du nombre de jours perdus à cause de la maladie. Shultz (2001) trouve une augmentation moyenne de 3,4 % de la scolarisation pour tous les élèves des niveaux 1 à 8 ; l'augmentation était la plus forte parmi les filles ayant obtenu le niveau 6, avec 14,8 %. En partie parce qu'il a été montré que le programme est une réussite, il fut par conséquent maintenu lorsque le gouvernement mexicain a changé de mains : en 2000, il touchait 2,6 millions de familles, 10 % des familles du Mexique, et avait un budget de 800 millions de dollars US, soit 0,2 % du PIB (Gertler et Boyce, 2001). Il fut ensuite étendu aux communautés urbaines et, avec le support de la Banque Mondiale, des programmes similaires sont en train d'être mis en place dans plusieurs pays voisins d'Amérique Latine. Les officiels mexicains ont transformé une contrainte budgétaire en une opportunité et ont fait de l'évaluation la pierre angulaire de l'extension qui a suivi. Ils furent récompensés par l'expansion du programme et par l'immense visibilité qu'il a acquise.

L'expansion de projets existants

Il est parfois possible d'évaluer l'impact de programmes qui ont déjà montré leur potentiel pour être mis en place à grande échelle. Au contraire des projets pilotes, on peut alors être sûr que le programme puisse être mis en place à grande échelle. Cela facilite également l'évaluation du programme sur plusieurs sites en même temps et évacue donc certaines inquiétudes concernant la validité externe. Une occasion naturelle pour une telle évaluation se présente lorsque le programme est prêt à être étendu et que l'expansion peut se dérouler dans un ordre aléatoire. L'évaluation d'un programme d'éducation de rattrapage par Banerjee, Cole, Duflo et Linden (2003) est un exemple de cette approche. Le programme a été exécuté par Pratham, une ONG indienne, qui l'avait mis en place en 1994. Pratham touche désormais plus de 161.000 enfants dans 20 villes. Le programme d'éducation de rattrapage embauche une jeune femme issue de la communauté des enfants pour fournir des cours de rattrapage dans les écoles publiques à des enfants qui ont atteint le grade 2, 3 ou 4 sans avoir maîtrisé les compétences

basiques du grade 1. Les enfants qui sont identifiés comme étant en retard sont retirés de la salle de classe habituelle pendant deux heures par jour afin de recevoir cette instruction. Pratham voulait évaluer l'impact de ce programme, une de ses interventions phares en même temps qu'elle cherchait à l'étendre. L'expansion dans une nouvelle ville, Vadodara, a donné une opportunité de conduire une évaluation par assignation aléatoire. Lors de la première année (1999-2000), le programme fut étendu à 49 (sélectionnées au hasard) des 123 écoles publiques de Vadodara. En 2000-2001, le programme fut étendu à toutes les écoles mais la moitié des écoles eut un professeur de rattrapage pour le grade 3 et l'autre moitié pour le grade 4. Les élèves du grade 3 dans les écoles qui ont eu le programme au grade 4 servent de groupe de comparaison pour les élèves du grade 3 dans les écoles qui ont eu le programme au grade 3. Dans le même temps, une intervention similaire fut conduite dans un district de Mumbai, où la moitié des écoles eut les professeurs de rattrapage dans le grade 2 et l'autre moitié dans le grade 3. Le programme fut poursuivi une année supplémentaire, les écoles échangeant les groupes. Le programme est donc conduit pour plusieurs grades, dans deux villes et sans aucune école se sentant lésée en ressources relativement aux autres puisque toutes les écoles ont bénéficié du programme. Après deux ans, le programme avait accru le résultat moyen des tests de 0,39 écarts-types (ce qui représente une augmentation de 3,2 points sur 100 possibles – la moyenne du groupe de contrôle était de 32,4 points), et un impact encore plus fort sur les résultats des tests des enfants qui avaient de faibles scores initialement (une augmentation de 3,7 points, ou 0,6 écarts-types, sur une base de 10,8 points). L'impact du programme augmente avec le temps mais il est très similaire d'une ville à l'autre ou selon le sexe des enfants. Embaucher des professeurs de rattrapage dans la communauté s'avère être 10 fois plus efficace en termes de coûts que l'embauche de nouveaux professeurs. On peut donc être relativement confiant en recommandant de faire passer ce programme au niveau supérieur, du moins en Inde, sur la base de ces estimations, à partir du moment où le programme fut poursuivi sur une période de temps, qu'il fut évalué dans deux contextes très différents et qu'il a montré sa capacité à être étendu à grande échelle.

L'assignation aléatoire induite par le programme

Dans quelques cas, des considérations d'équité et de transparence font de l'assignation aléatoire le meilleur moyen de choisir les bénéficiaires du programme. De tels programmes sont des candidats naturels

pour l'évaluation, puisque l'exercice d'évaluation ne requiert aucune modification de la conception du programme.

L'allocation à des écoles particulières est souvent faite par une loterie lorsque certaines écoles sont sur-demandées. Dans certains systèmes scolaires des États-Unis, les élèves ont l'option de postuler à des « écoles aimants » ou des écoles avec des programmes spéciaux, et l'admission est souvent attribuée par une loterie. Cullen, Jacob et Levitt (2002) utilisent ce dispositif pour évaluer l'impact du choix de l'école dans le système scolaire de Chicago en comparant les gagnants et les perdants de la loterie. Puisque chaque école mène sa propre loterie, leur article profite en effet de 1000 loteries différentes ! Ils trouvent que les gagnants d'une loterie ont moins de chances d'aller à l'école de leur quartier que les perdants mais ont plus de chances de rester dans le système scolaire de Chicago. Cependant, leur performance subséquente est en fait moins bonne que celle des perdants de la loterie. Ceci est un contraste important avec ce qui aurait été attendu et avec ce qu'une comparaison « naïve » aurait trouvé : les résultats des enfants qui allaient à l'école de leur choix sont en effet meilleurs que ceux qui n'ont pas choisi, mais cela reflète le fait que les enfants qui ont décidé de changer d'école étaient très motivés.

Les programmes de coupons constituent un autre exemple de programmes qui mettent souvent en place un dispositif de loterie : le gouvernement alloue seulement un budget limité au programme, le programme est sur-demandé et une loterie est utilisée pour choisir les bénéficiaires. Angrist *et al.* (2002) ont évalué un programme dans lequel des bons pour des écoles privées étaient alloués par une loterie à cause de la limitation du budget du programme. Les coupons étaient renouvelables sous condition de performance académique satisfaisante. Ils comparent les gagnants et les perdants de la loterie. Les gagnants de la loterie avaient 15-20 % de chances de plus d'aller dans une école privée, 10 % de chances de plus de réussir le 8^{e} grade et ont obtenu 0.2 écarts-types de plus lors de tests standardisés équivalents à un cursus complet. Les gagnants avaient substantiellement plus de chances d'obtenir leur baccalauréat et ont obtenu de meilleurs résultats lors des examens d'obtention du baccalauréat/d'entrée à l'université. Les bénéfices de ce programme pour les participants dépassaient clairement le coût, qui était similaire au coût de la fourniture d'une place dans une école publique.

Lorsque des politiques nationales incluent quelque aspect d'assignation aléatoire, cela fournit une opportunité d'évaluer une politique qui a déjà été étendue en plusieurs lieux. La connaissance acquise de cette

expérience peut être utilisée pour informer les décisions politiques d'étendre la politique dans le pays, de continuer le programme ou de l'étendre dans d'autres pays. Cependant, parce que l'assignation aléatoire fait partie de la conception du programme, au lieu d'être une tentative délibérée de rendre l'évaluation possible, les données qui rendent l'évaluation possible ne sont pas toujours disponibles. Les agences internationales peuvent jouer deux rôles clés à cet égard : premièrement, elles peuvent organiser et financer les efforts de collecte des données limitées; deuxièmement, elles peuvent encourager les gouvernements et les offices statistiques à relier les sources de données existantes qui peuvent être utilisées pour évaluer les expérimentations. Les quotas pour les femmes et les minorités dans le gouvernement décentralisé (le système Panchayat) en Inde sont un exemple intéressant. En 1993, le 73^{e} amendement à la Constitution de l'Inde exigeait des États qu'ils mettent en place un système Panchayat à trois niveaux (village, bloc et district), directement élu par le peuple, pour l'administration des biens publics locaux. Les élections doivent avoir lieu tous les cinq ans et les conseils Panchayat ont la latitude de décider comment allouer les dépenses d'infrastructures locales. L'amendement exigeait également qu'un tiers de tous les postes (de membres et de présidents du conseil) soit réservé aux femmes et qu'une part égale à la représentation des minorités désavantagées (castes et tribus prévues) soit réservée à ces minorités. Pour éviter toute manipulation possible, la loi stipulait que le poste réservé soit alloué au hasard. Chattopadhyay et Duflo (2001) ont évalué l'impact au Bengale Occidental de la réservation des sièges aux femmes. Ils ont collecté des données dans 465 villages, 165 conseils et un district et ils ont trouvé que les femmes tendaient à allouer plus de ressources à l'eau potable et aux routes et moins à l'éducation. Ceci correspond aux priorités exprimées par les hommes et les femmes au travers de leurs réclamations aux autorités du Panchayat. Avant de terminer une seconde esquisse de cet article (Chattopadhyay et Duflo 2003), ils ont collecté les mêmes données dans un district pauvre du Rajasthan, Udaipur. Ils ont trouvé que, là-bas, les femmes investissent plus dans l'eau potable et moins dans les routes, et que cela correspond encore une fois à l'ordre des réclamations exprimées par les hommes et les femmes. Ces résultats furent obtenus dans deux districts très différents avec des histoires différentes (le Bengale Occidental avait eu un Panchayat depuis 1978 alors que le Rajasthan n'en avait pas jusqu'en 1995; le Rajasthan est aussi un des États indiens avec une alphabétisation des femmes particulièrement faible), ce qui suggère que le sexe des décideurs politiques compte à la fois dans les systèmes politiques plus et moins développés. En outre, cela fournit une preuve indirecte (mais puissante) que les officiels

locaux élus ont vraiment du pouvoir, même dans des systèmes relativement « jeunes ». Ils ont aussi évalué l'impact de la réservation aux castes prévues et ont trouvé qu'une part plus importante des biens était attribuée aux hameaux des castes prévues quand le chef du Panchayat vient d'une caste prévue.

En principe, les données pour évaluer l'impact de cette expérimentation à plus grande échelle existent : les données de recensement au niveau des villages sont disponibles pour 1991 et vont devenir disponibles pour 2001. L'Organisation d'Enquêtes sur Echantillon National (NSSO) mène des enquêtes de consommation et d'emploi à grande échelle tous les cinq ans, avec des données détaillées sur les résultats. Cependant, les barrières administratives rendent ces données très difficiles à utiliser dans le but d'évaluer ce programme : le recensement ne contient aucune information sur le Panchayat auquel un village appartient. L'information sur la réservation et la composition d'un Panchayat n'est pas centralisée, même au niveau des États (elle est disponible uniquement au niveau du district). De même, la NSS ne contient aucune information sur le Panchayat. Il s'agit d'un exemple où, pour un coût relativement faible, il serait possible de rendre disponible cette information utile pour évaluer un programme très étendu. Cela requiert une coordination de différentes personnes et différentes agences, une tâche à accomplir pour laquelle les organisations internationales seraient bien placées.

3 LES LEÇONS

Les évaluations décrites dans la section 2 offrent à la fois des leçons réelles et méthodologiques. Dans ce qui suit, nous passons en revue quelques-unes des leçons méthodologiques qui peuvent être tirées des exemples discutés dans la section 2.

Les résultats des évaluations par assignation aléatoire peuvent être assez différents de ceux tirés d'évaluations rétrospectives

Lorsque l'évaluation n'est pas planifiée *ex ante*, afin d'évaluer l'impact d'un programme, les chercheurs doivent avoir recours à des comparaisons avant et après (quand une ligne de base a été menée), ou à des

comparaisons entre les bénéficiaires et les communautés qui, pour une raison quelconque, n'ont pas été exposés au programme. Lorsque les raisons pour lesquelles certaines personnes ont été exposées au programme et d'autres pas ne sont pas connues (ou pire, quand elles sont connues pour introduire probablement un biais de sélection), ces comparaisons ont des chances d'être biaisées. La collecte de données est souvent aussi chère que pour l'évaluation par assignation aléatoire mais les inférences sont biaisées. Comme nous l'avons affirmé plus haut, le contrôle pour des différences observables entre les groupes de traitement et de contrôle (au travers d'une analyse de régression ou au travers d'appariement par score de propension) sera correct pour le biais seulement si il est connu avec certitude que les bénéficiaires et les non bénéficiaires sont comparables conditionnellement à ces caractéristiques. Il est peu probable que ce soit vrai à moins que le programme ne fût alloué de manière aléatoire conditionnellement à ces caractéristiques. En particulier, un agent du projet tentant d'allouer de manière optimale un programme a typiquement plus d'information qu'un chercheur et va (et devrait) s'en servir en allouant les ressources.

Ces préoccupations ont d'importantes implications pratiques. Les études comparant les estimations expérimentales et non expérimentales avec les mêmes données montrent que les résultats de l'évaluation par assignation aléatoire peuvent être assez différents de ceux tirés d'évaluation sans assignation aléatoire. Dans une analyse célèbre des programmes de formation professionnelle, LaLonde (1986) a trouvé que de nombreuses procédures économétriques et groupes de comparaison utilisés dans les évaluations des programmes ne donnaient pas des estimations correctes ou précises et que de telles estimations économétriques diffèrent souvent de manière significative des résultats expérimentaux. Les études comparatives qui estiment l'impact d'un programme en utilisant des méthodes expérimentales et ré-estiment ensuite l'impact en utilisant une ou plusieurs méthodes non expérimentales différentes suggèrent que le biais de variable omise est un problème significatif au-delà des simples exemples mentionnés ici. Bien que nous n'ayons pas connaissance d'une quelconque revue systématique des études dans les pays en développement, une étude récente des pays développés suggère que le biais de variable omise est un problème majeur quand des méthodes non expérimentales sont utilisées (Glazerman *et al.*, 2002). Cette étude examine à la fois les études expérimentales et non expérimentales dans le contexte du bien-être, de la formation professionnelle et des programmes de services d'emploi et a trouvé que les estimateurs non expérimentaux produisent souvent des résultats nettement différents de ceux des évaluations par assignation aléatoire, que le biais estimé est

souvent important et qu'aucune stratégie ne semble bien se comporter de manière régulière[3].

Nous n'avons pas connaissance de meta-analyses systématiques pour les pays en développement, mais il y a de nombreux exemples où les évaluations par assignation aléatoire peuvent être comparées avec des évaluations prospectives utilisant la même base de données. Glewwe *et al.* (2003) ont comparé des analyses rétrospectives et prospectives de l'effet des tableaux de classes sur les résultats des tests. Les estimations rétrospectives utilisant de simples régressions en MCO suggèrent que les tableaux accroissent les résultats des tests jusqu'à 20 % d'un écart-type, résultat robuste à l'inclusion de variables de contrôle; les estimations de différences dans les différences suggèrent un effet plus faible d'environ 5 % d'un écart-type, un effet qui est encore significatif bien que parfois seulement au seuil de 10 %. À l'opposé, les estimations prospectives fondées sur les évaluations par assignation aléatoire ne fournissent aucune preuve que les tableaux de classe augmentent les résultats des tests. Ces résultats suggèrent que l'utilisation de données rétrospectives pour comparer des scores de test surestime sérieusement l'efficacité des tableaux. Une approche par différence dans les différences réduit mais n'élimine pas le problème et, de plus, il n'est pas clair qu'une telle approche par différence dans les différences puisse être applicable de manière générale. Ces exemples suggèrent que les estimations MCO sont biaisées vers le haut plutôt que vers le bas. C'est plausible, puisque dans un pays pauvre avec un rôle local substantiel dans l'éducation, les intrants sont probablement corrélés avec des caractéristiques favorables inobservées de la communauté. Si la direction du biais de variable omise était similaire dans les autres analyses rétrospectives des intrants d'éducation dans les pays en développement, les effets des intrants devraient être encore plus modestes que ce que les études rétrospectives suggèrent. Certains des résultats sont plus encourageants : par exemple, Buddlemeyer et Skoufias (2003) ont utilisé les résultats de l'évaluation par assignation aléatoire comme un point de repère pour examiner la performance du modèle de discontinuité des régressions afin d'évaluer l'impact du programme PROGRESA sur la santé et l'assiduité scolaire des enfants. Les chercheurs trouvent que la performance du modèle de discontinuité des régressions est remarquablement bonne dans ce cas :

3 Une étude récente non incluse dans l'analyse de Glazerman, Levy et Meyers (2002) est celle de Buddlemeyer et Skoufias (2003). Buddlemeyer et Skoufias utilisent les résultats de l'évaluation par assignation aléatoire comme un point de repère pour examiner la performance du modèle de discontinuité des régressions pour évaluer l'impact du programme PROGRESA sur la santé et l'assiduité scolaire des enfants et trouvent que la performance du modèle de discontinuité des régressions est bonne dans ce cas.

les estimations d'impact avec méthode quasi-expérimentale étaient en accord avec la preuve expérimentale dans dix des douze cas et les deux exceptions ont toutes deux eu lieu la première année du programme. De telles recherches peuvent fournir un conseil inestimable à propos de la validité et des biais potentiels des estimateurs quasi-expérimentaux.

De futures recherches dans ce sens seraient précieuses puisque de telles études comparatives peuvent aider à montrer le degré auquel les biais des estimations rétrospectives sont significatifs. Cependant, quand le groupe de comparaison pour les portions non expérimentales de ces études comparatives est décidé *ex post*, l'évaluateur devrait être capable de choisir parmi une variété de groupes de comparaison plausibles, certains devant avoir des résultats concordant avec ceux des estimations expérimentales et d'autres non (Comme cela est discuté dans ce qui suit, c'est également un problème pour les études rétrospectives au regard des problèmes avec les biais de publication). Des moyens possibles d'aborder ces préoccupations dans le futur incluent la conduite d'évaluations non expérimentales d'abord, avant que les résultats des évaluations par assignation aléatoire ne soient rendus publics, ou d'avoir des chercheurs pour mener des évaluations non expérimentales aveugles, sans la connaissance des résultats des évaluations par assignation aléatoire ou des autres études non expérimentales.

Les évaluations par assignation aléatoire sont souvent faisables

Comme nous l'avons noté dans l'introduction, les évaluations par assignation aléatoire ne sont pas adaptées pour tous les types de programmes. Elles sont adaptées aux programmes qui ciblent des individus ou des communautés et où les objectifs sont bien définis. Par exemple, l'efficacité de l'aide extérieure déboursée en tant que support au budget général ne peut pas être évaluée de cette manière. Il serait désirable, pour des motifs politiques ou d'efficacité, de débourser une fraction de l'aide sous cette forme, bien que cela serait extrêmement coûteux de distribuer toute l'aide extérieure sous la forme d'un support au budget général, précisément parce que cela ne laisse aucune place à une évaluation rigoureuse des projets. Cependant, comme l'exemple pris dans cet article le démontre, dans de nombreux cas, les évaluations par assignation aléatoire sont faisables. Le coût principal de l'évaluation est le coût de la collecte de données, et ce n'est pas plus coûteux que la collecte d'autres données quelconques. En fait, imposer une discipline sur les données à collecter (les résultats d'intérêt sont définis ex ante

et n'évoluent pas quand le programme ne parvient pas à les affecter) devrait réduire le coût de la collecte de données, relativement à une situation où ce qui est mesuré n'est pas clair.

Des préoccupations d'économie politique rendent parfois difficile de ne pas mettre en place le programme sur l'ensemble de la population, particulièrement quand son succès a déjà été démontré (par exemple, « opportunidades », la version urbaine de PROGRESA, ne débutera pas avec une évaluation par assignation aléatoire, à cause de la forte opposition au fait d'en retarder l'accès à certaines personnes). Cette objection peut être attaquée à plusieurs niveaux. Premièrement, l'opposition à l'assignation aléatoire a moins de chances de voir le jour dans un environnement où elle jouit d'un fort support, en particulier si une règle prescrit qu'une évaluation est nécessaire avant la mise en place à grande échelle. Deuxièmement, si, comme nous l'avons affirmé plus haut, les évaluations ne sont pas financées par des prêts mais par des subventions, cela faciliterait la tâche de convaincre les partenaires de son utilité, en particulier si cela peut permettre au pays d'étendre un programme. Un exemple d'un tel partenariat explicite est une étude de l'efficacité de l'éducation sur le VIH/SIDA, actuellement conduite au Kenya (Duflo, Dupas, Kremer et Sinei 2003). Avec le support de l'UNICEF, le gouvernement du Kenya a rassemblé un programme d'enseignement-apprentissage pour l'éducation sur le VIH/SIDA. Par manque de fonds, la couverture du programme est restée très partielle. Le Partenariat pour le Développement de l'Enfant, avec des subventions de la Banque Mondiale, finance actuellement une évaluation par assignation aléatoire du programme d'enseignement-apprentissage. ICS, une ONG hollandaise, organise des sessions d'entraînement en compagnie de facilitateurs du gouvernement kenyan. L'évaluation a donné la possibilité d'étendre l'entraînement à 540 enseignants de 160 écoles, ce qui n'aurait pas été possible sinon. L'assignation aléatoire ne fut pas un motif de rejet du programme par les autorités kenyanes. Au contraire, lors d'une conférence organisée pour le lancement du programme, les officiels kenyans ont explicitement apprécié l'opportunité que l'évaluation leur donnait d'être à la pointe des efforts pour faire avancer la connaissance sur la question. L'exemple de PROGRESA a montré que les officiels du gouvernement ont reconnu la valeur de l'évaluation par assignation aléatoire et sont en fait préparés à payer pour cela. La réponse très favorable à PROGRESA et l'endossement subséquent des résultats par la Banque Mondiale auront certainement un impact sur l'opinion des autres gouvernements concernant les expérimentations. Plusieurs exemples de ce type pourraient faire beaucoup pour changer la culture.

Les ONG sont très aptes à mener les évaluations par assignation aléatoire mais vont avoir besoin d'une assistance technique (par exemple, de la part des universitaires) et d'un financement extérieur.

Les gouvernements ne sont pas les seuls vecteurs par lesquels les évaluations par assignation aléatoire peuvent être organisées. À vrai dire, l'évidence présentée dans cet article suggère qu'un modèle possible est celui de l'évaluation des projets des ONG. À la différence des gouvernements, on n'attend pas des ONG qu'elles servent la population entière. Même les petites ONG peuvent affecter substantiellement les budgets dans les pays en développement. Étant donné qu'il existe beaucoup d'ONG et qu'elles cherchent fréquemment de nouveaux projets, il est souvent relativement correct de trouver des ONG voulant mener des évaluations par assignation aléatoire : les réticences sont plus souvent logistiques que philosophiques.

Par exemple, une série d'études récentes conduites au Kenya a été menée au travers d'une collaboration avec l'ONG kenyane Internationaal Christelijk Steufonds (ICS) Africa : ICS était vivement intéressée par l'utilisation d'évaluations par assignation aléatoire pour voir l'impact qu'avaient ses programmes, ainsi que pour partager des résultats d'évaluation crédibles avec les autres dépositaires et les décideurs politiques. Un second exemple est celui de la collaboration entre l'ONG indienne Pratham et les chercheurs du MIT, qui a conduit à des évaluations des programmes d'éducation de rattrapage et d'enseignement assisté par ordinateur (Banerjee *et al.*, 2003). Cette collaboration fut initiée quand Pratham cherchait des partenaires pour évaluer ses programmes; Pratham comprit la valeur de l'assignation aléatoire et fut capable de transmettre l'importance de telles évaluations aux instituteurs impliqués dans le projet.

Cependant, alors que les ONG sont bien placées pour mener des évaluations par assignation aléatoire, il est moins raisonnable d'en attendre qu'elles financent ces évaluations. Les évaluations des programmes de suppression des vers ont été rendues possibles par un soutien financier de la Banque Mondiale, du Partenariat pour le Développement de l'Enfant, de l'Institut National de la Santé américain (NIH) et de la Fondation MacArthur. Dans le cas des programmes éducatifs indiens, Pratham fut capable de trouver une firme de parrainage; la deuxième banque d'Inde, la banque ICICI, était vivement intéressée par l'évaluation de l'impact du programme et aida à financer une partie de l'évaluation. En général, étant donné que les estimations précises des effets d'un programme sont des biens publics

internationaux, les évaluations par assignation aléatoire devraient être financées internationalement.

Les coûts peuvent être réduits et la comparabilité accrue en menant une série d'évaluations dans la même zone

Une fois que les membres de l'équipe sont entraînés, ils peuvent travailler sur de multiples projets. Puisque la collecte de données est l'élément le plus coûteux de ces évaluations, croiser l'échantillon peut également réduire fortement les coûts. Par exemple, beaucoup de programmes cherchant à accroître la participation scolaire furent mis en place dans la même zone et par la même organisation. Des programmes d'incitations au professeur (Glewwe *et al.*, 2003) et de manuels scolaires furent évalués dans les 100 mêmes écoles du Kenya Occidental : un groupe n'avait que les manuels, un avait les manuels et les incitations, un n'avait que les incitations, et un n'en avait aucun des deux. L'effet du programme d'incitation devrait donc être interprété comme l'effet d'un programme d'incitation conditionnel au fait que la moitié des écoles avait des manuels supplémentaires. De même, à Vadodara, Pratham avait mis en place un programme d'enseignement assisté par ordinateur dans les mêmes écoles que celles où le programme d'éducation de rattrapage évalué par Banerjee *et al.* (2003) avait été mis en place. Le programme avait été mis en place uniquement au grade 4. La moitié des écoles qui avaient le programme d'éducation de rattrapage au grade 4 a eu le programme d'enseignement assisté par ordinateur et la moitié des écoles qui n'avaient pas le programme d'éducation de rattrapage a eu le programme d'enseignement assisté par ordinateur. Les résultats préliminaires suggèrent que l'effet sur les maths est comparable à l'effet de l'éducation de rattrapage mais le coût est bien plus faible. Même en maintenant constant le budget de l'évaluation du processus, une réallocation d'une partie de l'argent qui est actuellement dépensé en évaluation non convaincante irait probablement *vers* le financement du même nombre d'évaluations par assignation aléatoire. Même si les évaluations par assignation aléatoire s'avèrent être plus chères, le coût est probablement insignifiant en comparaison de la quantité d'argent économisée en évitant l'expansion de programmes inefficaces. Cela suggère que l'évaluation aléatoire devrait être financée par les organisations internationales, un point sur lequel nous reviendrons dans ce qui suit.

Cette technique doit prendre en compte les interactions possibles entre les programmes (qui peuvent être estimées si l'échantillon est

suffisamment grand) et peut ne pas être appropriée si un programme rend les écoles atypiques. Mais elle présente l'avantage de rendre possible l'évaluation de l'efficacité-coût de différentes approches pour combattre le même problème. Par exemple, plusieurs évaluations menées en Inde dans le même ensemble d'écoles ont aidé à conclure que la suppression de vers était le moyen le plus efficace en termes de coût d'accroître la participation scolaire (comparé aux uniformes scolaires, aux repas scolaires et aux incitations aux professeurs) (Kremer, 2003), alors que les bourses attribuées au mérite sont le moyen le plus efficace en termes de coût pour accroître le score aux tests (Kremer *et al.* (2004)).

Le minutage de l'évaluation et de la mise en œuvre

Les évaluations prospectives prennent du temps : les études convaincantes durent souvent deux ou trois ans. Cela prend encore plus de temps d'obtenir l'impact de long terme du programme, qui peut être très important et différent de l'impact à court terme. Par exemple, Glewwe, Illias et Kremer (2003) suggèrent qu'un programme d'incitation aux enseignants provoque une augmentation des scores aux tests à court terme, mais aucun impact à long terme, ce qu'ils attribuent aux pratiques « d'enseigner pour le test ». Lorsque le programme cible des enfants mais cherche à affecter les résultats d'un adulte (ce qui est le cas pour la plupart des interventions dans l'éducation ou la santé), le délai entre le programme et les résultats peut devenir très long. Dans ces cas, il n'est pas possible d'attendre la réponse avant de décider de mettre en place ou non le programme.

Bien que cela soit une préoccupation réelle, cela ne doit pas empêcher la mise en place de l'évaluation sur la première cohorte à être exposée au programme : bien qu'il soit vrai que les décisions politiques vont devoir être prises dans l'intervalle, il est sûrement meilleur de connaître la réponse à un moment donné plutôt que jamais, ce qui serait le cas sans évaluation. De plus, il est souvent possible d'obtenir des résultats à court terme, ce qui devrait être utilisé pour obtenir une indication des chances du programme d'être efficace ou non, et devrait également guider la politique à court terme. Par exemple, dans le cas du programme d'entraînement des enseignants VIH/SIDA, une évaluation fut réalisée quelques semaines après que le programme n'ait débuté (et alors qu'il était toujours en cours) : les élèves des écoles où les enseignants avaient été entraînés les premiers furent interrogés sur la présence ou non du VIH/SIDA au programme scolaire dans leur école, et ils furent également soumis à un test de connaissances, d'attitude et

de pratique. Les résultats préliminaires suggèrent que le programme était bien efficace pour accroître les chances que le VIH/SIDA soit mentionné en classe et pour accroître la connaissance des élèves sur le VIH/SIDA et sur la prévention contre le VIH. Ces résultats pourraient être communiqués immédiatement aux décideurs politiques. Le premier résultat d'une évaluation peut aussi être combiné avec d'autres résultats ou avec la théorie pour fournir une estimation de l'impact final attendu du programme. Evidemment, il faut être très prudent concernant ces exercices et prudemment distinguer ce qui provient des résultats de l'évaluation et ce qui est le résultat d'hypothèses. On devrait mettre en place des programmes pour être capable de suivre les résultats à long terme, ce qui peut alors justifier ou invalider ces prédictions. Par exemple, Miguel et Kremer (2003) ont combiné leur estimation de l'impact du programme de suppression de vers sur la participation scolaire sur des estimations des rendements de l'éducation au Kenya pour fournir une estimation de l'impact de long terme sur la productivité des adultes, qu'ils ont utilisée pour construire leurs estimations coût-bénéfice. Ils continuent également à suivre les enfants exposés au médicament de suppression de vers pour estimer leur effet de long terme.

Enfin, retarder certaines dépenses vaudrait en fait la peine, étant donné que nous en connaissons si peu sur ce qui fonctionne et ce qui ne fonctionne pas, en particulier si cela peut nous donner une opportunité d'en apprendre davantage. Il est très déconcertant que nous n'en sachions pas plus sur ce qui fonctionne et ce qui ne fonctionne pas dans l'éducation, par exemple, après avoir passé tant d'années à financer des projets d'éducation. Sur cette échelle, le fait qu'une évaluation prenne deux ou trois ans (ou même bien plus pour obtenir des informations sur les résultats de long terme) semble être une période de temps très courte. Cela pourrait retarder quelques dépenses, mais cela accélèrera le processus d'apprentissage pour rendre ces dépenses utiles. La FDA exige une évaluation par assignation aléatoire des effets d'un médicament avant qu'il puisse être distribué. Occasionnellement, le délai qu'elle impose sur l'approbation de nouveaux médicaments a provoqué du ressentiment (le plus récemment, parmi les associations représentant les victimes du SIDA). Cependant, il y a peu de doutes sur le fait que les tests par assignation aléatoire ont joué un rôle clé dans le façonnement de la médecine moderne et qu'ils ont accéléré le développement de médicaments efficaces.

Les évaluations par assignation aléatoire ont un certain nombre de limites mais beaucoup de ces limites s'appliquent aussi aux autres techniques

De nombreuses limites des évaluations par assignation aléatoire s'appliquent aussi aux autres techniques. Dans cette sous-section, nous passons en revue quatre problèmes qui affectent à la fois les évaluations par assignation aléatoire et sans assignation aléatoire (le biais de sélection de l'échantillon, le biais d'attrition, les effets de débordement et les réponses comportementales) et nous affirmons que les méthodes avec assignation aléatoire permettent souvent une correction plus facile de ces limites que ne le font les méthodes sans assignation aléatoire.

Les problèmes de sélection de l'échantillon peuvent apparaître si des facteurs autres que l'attribution aléatoire influencent l'allocation du programme. Par exemple, des parents pourraient enlever leurs enfants d'une école sans le programme pour l'inscrire dans une école avec le programme. Inversement, des individus alloués à un groupe de traitement pourraient ne pas recevoir le traitement (par exemple, parce qu'ils décident de ne pas adopter le programme). Même si des méthodes avec assignation aléatoire ont été utilisées et que l'allocation du programme prévue était aléatoire, la véritable allocation pourrait ne pas l'être. Ce problème peut être traité au travers de méthodes « d'intention de traiter (ITT) » ou en utilisant la désignation aléatoire comme un instrument des variables pour la désignation réelle. Bien que l'attribution initiale ne garantisse pas dans ce cas que quelqu'un soit véritablement soit dans le groupe du programme ou dans celui de comparaison, dans la plupart des cas, il est au moins plus probable que quelqu'un soit dans le groupe du programme si il ou elle y était initialement alloué. Le chercheur peut donc comparer les résultats dans le groupe alloué initialement et augmenter la différence, en la divisant par la différence dans la probabilité de recevoir le traitement dans ces deux groupes, pour obtenir l'estimation de l'effet local moyen du traitement (Imbens et Angrist, 1994). Des méthodes telles que les estimations ITT permettent de traiter les problèmes de sélection assez facilement dans le contexte d'évaluations par assignation aléatoire, mais il est souvent beaucoup plus difficile de faire ces corrections dans le cas d'une analyse rétrospective.

Un second problème affectant à la fois les évaluations avec et sans assignation aléatoire est l'attrition différentielle dans les groupes de traitement et de comparaison : ceux qui participent au programme devraient avoir moins de chances de sortir de ou sinon d'abandonner l'échantillon que ceux qui n'y participent pas. Par exemple, le programme avec deux

enseignants analysé par Banerjee *et al.* (2001) a accru l'assiduité scolaire et a réduit les taux d'abandon. Cela signifie que quand un test a été administré dans les écoles, plus d'enfants étaient présents dans les écoles du programme que dans les écoles de comparaison. Si les enfants que le programme empêche d'abandonner sont les plus faibles de la classe, la comparaison entre les scores de test des enfants issus des écoles de traitement et celles de contrôle devrait être biaisée vers le bas. Des techniques statistiques peuvent être utilisées pour borner le biais potentiel mais l'idéal est de tenter de limiter l'attrition autant que possible. Par exemple, dans l'évaluation du programme d'éducation de rattrapage en Inde (Banerjee *et al.*, 2003) une tentative fut faite de dénicher *tous* les enfants et de leur administrer le test, même s'ils avaient abandonné l'école. Seuls les enfants qui étaient partis pour rentrer dans leur village d'origine ne furent pas testés. En conséquence, le taux d'attrition est resté relativement haut mais n'était pas différent entre les écoles de traitement et de comparaison – ce qui accrût la confiance dans les estimations.

Troisièmement, les programmes peuvent créer des effets de débordement sur des personnes qui n'ont elles-mêmes pas été traitées. Ces débordements peuvent être physiques, comme cela fut trouvé pour le programme kenyan de suppression de vers par Miguel et Kremer (2003, à paraître) quand la suppression de vers interfère avec la transmission de maladies et réduit alors l'infection par vers à la fois parmi les enfants des écoles du programme qui n'avaient pas reçu le médicament et parmi les écoles voisines. De tels effets de débordement pourraient aussi opérer à travers les prix, comme lorsque la fourniture de repas scolaires conduit les écoles locales concurrentes à réduire leurs droits d'inscription (Vermeersch, 2002).

Enfin, il peut aussi y avoir des effets d'apprentissage et d'imitation (Duflo et Saez, à paraître; Miguel et Kremer, 2003b).

Si de tels effets de débordements sont globaux (par exemple, du fait de changements dans les prix mondiaux), les impacts totaux du programme seront difficiles à identifier avec n'importe quelle méthodologie. Cependant, si de tels débordements sont locaux, alors l'assignation aléatoire au niveau des groupes peut permettre l'estimation de l'effet total du programme à l'intérieur des groupes et peut générer une variation suffisante dans la densité du traitement local pour mesurer les débordements entre les groupes. Par exemple, la solution dans le cas de l'étude de la suppression de vers fut de choisir *l'école* (plutôt que les élèves au sein d'une école) comme unité d'assignation aléatoire (Miguel et Kremer, 2003, à paraître) et de regarder le nombre d'écoles

de traitement et de comparaison dans le voisinage. Bien évidemment, cela exige une taille d'échantillon plus grande.

Un problème qui pourrait ne pas être traité aussi facilement est lié au fait que la fourniture d'intrants pourrait temporairement accroître le moral parmi les élèves et les enseignants, et par conséquent améliorer la performance. Alors que cela biaiserait les évaluations par assignation aléatoire, cela biaiserait également les estimations en effets fixes ou en différences dans les différences. Cependant, la gravité du problème en pratique n'est pas claire, attendu que nous savons que la sélection est une préoccupation grave.

En résumé, bien que l'évaluation aléatoire ne soit pas une stratégie à l'épreuve des balles, les biais potentiels sont bien connus et peuvent souvent être corrigés. Cela la pose à l'opposé des biais de la plupart des autres types d'études, où le biais dû à des attributions non aléatoires du traitement ne peut pas être signé ni estimé.

Le biais de publication apparaît comme étant substantiel dans les études rétrospectives; les évaluations par assignation aléatoire peuvent aider à traiter ces problèmes de biais de publication mais on a besoin des institutions

Le biais de publication est un problème particulièrement important qui doit être traité. Des résultats positifs tendent naturellement à recevoir une grande quantité de publicité : les agences qui exécutent les programmes cherchent de la publicité pour leurs projets ayant réussi et les universitaires sont beaucoup plus intéressés par et capables de publier des résultats positifs au lieu de résultats modestes ou insignifiants. Cependant, de nombreux programmes échouent clairement et le biais de publication sera substantiel si les résultats positifs ont beaucoup plus de chances d'être publiés. L'évidence disponible suggère que le problème de biais de publication est grave (DeLong et Lang, 1992) et particulièrement significatif pour les études qui emploient des méthodes non expérimentales.

Le biais de publication a des chances d'être un problème particulier concernant les études rétrospectives. *Ex post*, les chercheurs ou les évaluateurs définissent leur propre groupe de comparaison et devraient alors être capables de choisir une variété de groupes de comparaison plausibles; en particulier, les chercheurs obtenant des résultats négatifs avec des techniques rétrospectives vont probablement tenter des approches différentes, ou ne pas publier. Dans le cas « d'expérimentations naturelles » et des estimations en variable instrumentale, le biais de publi-

cation peut en fait plus que compenser la réduction du biais causée par l'utilisation d'une variable instrumentale parce que ces estimations tendent à avoir des écarts-types plus grands et parce que les chercheurs cherchant des résultats significatifs vont sélectionner uniquement les grandes estimations. Par exemple, Ashenfelter *et al.* (1999) montrent qu'il y a une preuve très solide d'un biais de publication dans les estimations des rendements de l'éducation fondées sur des variables instrumentales : en moyenne, les estimations avec de plus grands écarts-types tendent elles aussi à être plus grandes. Cela compte pour une grande partie du résultat souvent cité selon lequel les estimations instrumentales des rendements de l'éducation sont plus élevées que les estimations en moindres carrés ordinaires.

À l'opposé, les évaluations par assignation aléatoire s'engagent à l'avance avec un groupe de comparaison particulier : une fois que le travail de mener une évaluation par assignation aléatoire prospective est effectué, les résultats sont d'ordinaire documentés et publiés même si les résultats suggèrent des effets plutôt modestes voire même aucun effet du tout.

Comme nous en discuterons plus loin, il est important de mettre en place des institutions qui assurent la diffusion des résultats négatifs. Un tel système est déjà en place pour les résultats des essais médicaux et créer un système similaire pour documenter les évaluations de programmes sociaux aiderait à réduire le problème du biais de publication. Au-delà du fait de permettre une représentation plus claire des interventions qui ont fonctionné et de celles qui n'ont pas fonctionné, ce type d'institution fournirait le niveau de transparence nécessaire pour faire en sorte que les revues de la littérature systématiques soient moins biaisées dans leurs conclusions sur l'efficacité de politiques et de programmes particuliers.

Bien que n'importe quelle évaluation par assignation aléatoire donnée soit menée dans un cadre spécifique avec des circonstances uniques, les évaluations par assignation aléatoire peuvent éclairer des problèmes généraux

Sans une théorie expliquant pourquoi un programme a l'effet qu'il a, la généralisation à partir d'une évaluation par assignation aléatoire bien exécutée peut être injustifiée. Mais des problèmes similaires de capacité de généralisation surviennent quelle que soit la technique d'évaluation utilisée. Un moyen d'en apprendre à propos de la capacité de généralisation est d'encourager des répliques adaptées d'évaluations

par assignation aléatoire dans des domaines d'intérêt clés sous plusieurs cadres différents. Il sera toujours possible qu'un programme qui a échoué dans un contexte aurait pu réussir dans un autre, mais des répliques adaptées, guidées par une théorie sur les raisons pour lesquelles le programme était efficace, seront un grand pas vers la réduction de cette préoccupation. C'est un domaine où les organisations internationales, qui sont déjà présentes dans la plupart des pays, peuvent jouer un rôle clé. Une telle opportunité fut saisie lors de la mise en place de répliques adaptées de PROGRESA dans d'autres pays d'Amérique Latine. Encouragée par le succès de PROGRESA au Mexique, la Banque Mondiale a encouragé (et financé) les voisins du Mexique à adopter des programmes similaires. Certains de ces programmes ont inclus des évaluations par assignation aléatoire (par exemple le programme Programa de Asignacion Familiar (PRAF) au Honduras) et sont actuellement en cours d'évaluation, avec quelques variantes qui nous aideront à mieux comprendre l'impact des différentes règles du programme.

Les résultats de la première phase d'un projet peuvent souvent être difficiles à interpréter à cause de circonstances uniques à la première phase : un projet peut avoir échoué en résultat de problèmes de mise en œuvre qui pourraient être évités dans les phases suivantes du projet ; ou un projet peut avoir réussi parce qu'il a reçu plus de ressources qu'un projet dans une situation plus réaliste ou un contexte moins favorable. Même si le choix des groupes de comparaison et de traitement assure la validité interne des estimations, toute méthode d'évaluation est sujette à des problèmes de validité externe dus aux circonstances spécifiques de la mise en œuvre. De fait, les résultats peuvent ne pas être généralisables à d'autres contextes.

Un problème spécifique aux évaluations par assignation aléatoire est le fait que des membres du groupe de traitement ou de comparaison peuvent potentiellement changer leur comportement, pas du fait de l'intervention, mais simplement du fait qu'ils pourraient savoir qu'ils font partie de l'évaluation par assignation aléatoire. Par exemple, la fourniture d'intrants pourrait temporairement accroître le moral au sein des bénéficiaires et cela pourrait améliorer la performance. Bien évidemment, dans la mesure où les deux groupes changent leur comportement de la même manière, cela ne mènera pas à un biais. Il est aussi peut-être moins probable que cela se produise sur une longue période et que cela se produise immédiatement après l'introduction de l'intervention. Certaines conceptions expérimentales peuvent minimiser le risque de tels effets. Par exemple, dans le programme d'éducation

de rattrapage de Pratham analysé par Banerjee *et al.* (2003), *toutes* les écoles ont reçu le programme, mais pas tous les grades. Il est cependant important de tenter d'établir si ces effets sont présents. Dans sa ré-analyse des données du projet STAR, Krueger (1999) exploite la variation dans les tailles de classe au sein du groupe de contrôle occasionnée par le départ d'enfants au cours de l'année pour obtenir une seconde estimation de l'effet de la taille de classe, qui n'est pas, par définition, contaminé par les effets John Henry ou Hawthorne, puisque tous les enseignants dans cet échantillon appartiennent au groupe de contrôle. Il ne trouve aucune différence entre les estimations obtenues par ces deux méthodes.

Les effets du traitement peuvent aussi être affectés par l'échelle du programme. Par exemple, le programme de coupons colombien analysé par Angrist *et al.* (2002) que nous avons décrit plus haut fut mis en œuvre sur une base pilote avec un petit échantillon mais le reste du système scolaire resta inchangé (en particulier, le nombre d'élèves affectés était trop petit pour avoir un impact sur la composition des écoles publiques et privées). Si ce programme devait être mis en œuvre à grande échelle, il pourrait affecter le fonctionnement du système scolaire et avoir alors un impact différent (Hsieh et Urquiola 2002). Plus généralement, les effets de traitement en « équilibre partiel » peuvent être différents des effets de traitement en « équilibre général » (Heckman, Lochner et Taber 1998). Pour traiter ces problèmes, nous avons besoin d'une évaluation avec assignation aléatoire exécutée au niveau de « l'économie ». Cela peut être possible pour des programmes tels que le programme de coupons, où les effets d'équilibre général vont plausiblement prendre place au niveau de la communauté, et où les communautés peuvent être affectées ou non par le programme de manière aléatoire. Mais je n'ai pas connaissance d'une évaluation de ce type.

Un moyen de traiter les questions concernant la validité externe de n'importe quelle étude particulière, que ce soit une évaluation par assignation aléatoire ou non, est de mettre en œuvre des répliques adaptées de programmes ayant réussi (et potentiellement n'ayant pas réussi) dans différents contextes. De telles répliques adaptées ont deux avantages : premièrement, dans le processus de « transplantation » d'un programme, les circonstances vont changer et les programmes robustes montreront leur efficacité en survivant à ces changements ; deuxièmement, obtenir plusieurs estimations dans des contextes différents fournira quelques conseils concernant le fait que le programme ait des impacts notablement différents sur des groupes différents. Deux études portant sur des

interventions de santé basées à l'école fournissent une bonne illustration. La première étude (Miguel et Kremer 2003) a évalué un programme de traitement de masse semestriel basé à l'école utilisant un médicament de suppression de vers au Kenya, où la prévalence de vers intestinaux parmi les enfants est très élevée. Soixante-quinze écoles furent introduites graduellement dans le programme selon un ordre aléatoire. La santé et la participation scolaire se sont améliorées, pas uniquement dans les écoles du programme, mais aussi dans les écoles proches, du fait d'une transmission réduite de la maladie. L'absentéisme dans les écoles de traitement était de 25 % (ou 7 points de pourcentage) moindre que dans les écoles de comparaison. En incluant cet effet de débordement, le programme accrût la scolarisation de 0,15 ans par personne traitée. Combinées avec des estimations sur le taux de rendement de la scolarisation, les estimations suggèrent des taux de rendement de l'intervention de suppression de vers extrêmement élevés : les auteurs estiment que la suppression de vers accroît la valeur actuelle nette des salaires de plus de 30$ par enfant traité à un coût de 0,49$ seulement. Un des auteurs a alors décidé d'examiner si ces résultats se généralisaient parmi les enfants en maternelle en Inde urbaine (Bobonis, Miguel et Sharma 2002). L'étude de base révéla que, bien que l'infection au vers soit présente, les niveaux d'infection étaient substantiellement plus faibles qu'au Kenya (en Inde, « seulement » 27 % des enfants souffrent d'une forme quelconque d'infection aux vers). Cependant, 70 % des enfants souffrent d'anémie modérée à sévère. Le programme fut donc modifié pour inclure un supplément alimentaire en fer. Le programme fut administré à travers un réseau d'écoles maternelles en Inde urbaine. Après un an de traitement, ils ont trouvé une réduction d'environ 50 % de l'anémie modérée à sévère, d'importants gains de poids et une réduction de 7 % de l'absentéisme parmi les enfants âgés de 4 à 6 ans (mais par pour les enfants plus jeunes). Les résultats de l'évaluation précédente étaient donc à tout prendre justifiés[4].

La réplique de la phase initiale d'une étude dans un nouveau contexte n'implique pas de retarder la mise en œuvre à grande échelle du programme si c'est justifié sur la base de la connaissance existante. Le plus souvent, l'introduction du programme ne peut malgré tout procéder que par étapes et l'évaluation requiert uniquement que les participants soient introduits graduellement dans le programme dans un ordre aléatoire. De plus, de telles répliques peuvent être utilisées pour vérifier si les effets du programme au sein des échantillons varient avec la

4 Pour être tout à fait précis, on aurait besoin d'une analyse coût-bénéfice complète sur les deux programmes pour voir si la même amélioration en capital humain fut obtenue avec la même dépense. À ce niveau, l'article sur l'Inde n'a pas encore d'analyse coût-bénéfice.

covariance. Par exemple, supposons que l'effet d'un programme donné soit plus faible dans les écoles avec de bons enseignants; on pourrait réfléchir à la possibilité que l'effet soit plus faible dans un contexte différent avec des enseignants encore meilleurs.

On requiert des institutions qu'elles fournissent des incitations à mener de telles répliques et qu'elles agrègent les résultats pour obtenir une représentation cohérente de l'impact d'une approche particulière.

Cela vaut la peine de noter que la variation exogène créée par l'assignation aléatoire peut être utilisée pour aider à identifier un modèle structurel. Attanasio *et al.* (2001) et Berhman *et al.* (2002) sont deux exemples de l'utilisation de cet exercice en combinaison avec les données PROGRESA pour prédire les effets possibles d'une variation du calendrier des transferts. Par exemple, Attanasio *et al.* (2001) ont trouvé que la composante sélectionnée aléatoirement des données PROGRESA induisait une variation exogène extrêmement utile qui a aidé à l'identification d'un modèle structurel plus riche et plus flexible. Ces études reposent sur les hypothèses auxquelles chacun est libre d'adhérer ou non mais elles sont au moins libérées de *certaines* hypothèses par la présence de cette variation exogène.

Le point le plus général est le fait les évaluations par assignation aléatoire n'excluent pas l'utilisation d'une théorie ou d'hypothèses : en fait, elles génèrent des données et de la variation qui peuvent être utiles pour identifier certains aspects de ces théories.

Une théorie expliquant pourquoi un programme spécifique a des chances d'être efficace est nécessaire pour fournir quelques éclairages sur les éléments du programme et de son contexte qui furent les clés de son succès. De manière importante, la théorie aidera à déballer les composantes distinctes d'un programme et à discriminer entre les variantes qui ont des chances d'être importantes et les variantes qui ne le sont pas (Banerjee 2002). Par exemple, une analyse économique du programme PROGRESA suggère qu'il aurait pu être utile à cause de son impact sur le revenu, sur le pouvoir de négociation des femmes ou à cause de son effet sur les incitations. Les aspects du programme ayant le plus de chance d'être pertinents pour le succès du programme sont la taille du transfert, son destinataire et la conditionnalité qui y est attachée. Par opposition, la couleur du supplément alimentaire distribué aux familles, par exemple, n'est probablement pas importante. La réplique des programmes peut alors varier ces différents aspects pour déterminer lequel est le plus important. Cela suggère également que les programmes qui sont justifiés par un quelconque raisonnement théorique bien fondé devraient être évalués en priorité parce les conclusions de

l'évaluation ont alors plus de chances d'être généralisées. La théorie fournit quelques éclairages sur les programmes ayant des chances de fonctionner et, à son tour, l'évaluation de ces programmes forme un test de la prédiction de la théorie. Puisque les évaluations prospectives doivent être planifiées à l'avance, il est aussi souvent possible de concevoir les programmes pilotes de telle manière à ce qu'ils aident à répondre à une question spécifique ou à tester une théorie spécifique. Par exemple, Duflo (2003) fait un rapport d'une série d'évaluations par assignation aléatoire menées au Kenya avec Michael Kremer et Jonathan Robinson. Ils étaient motivés par la question générale : pourquoi y a t-il si peu de fermiers dans cette région du Kenya qui utilisent des fertilisants (environ 10 % d'entre eux seulement en utilisent) en dépit du fait qu'ils semblent être rentables et qu'ils sont largement utilisés dans les autres pays en développement ainsi que dans les autres régions du Kenya ? Dans un premier temps, ils ont mené une série d'essais sur les fermes détenues par les fermiers sélectionnés au hasard et ont confirmé que, en petite quantité, les fertilisants sont extrêmement rentables (les taux de rendements dépassaient souvent les 100 %). Ils ont alors mené une série de programmes pour répondre aux questions suivantes : les fermiers apprennent-ils lorsqu'ils essaient les fertilisants pour eux-mêmes ? Ont-ils besoin d'informations sur les rendements ou sur la façon de les utiliser ? L'expérimentation doit-elle prendre place dans leur ferme ou peut-elle prendre place dans la ferme d'un voisin ? Apprennent-ils de leurs amis ? Pour y répondre, ils ont mis en œuvre plusieurs programmes : premièrement, ils ont sélectionné au hasard des fermiers pour participer aux essais de terrain et ont suivi leur adoption ultérieure, de même que celle du groupe de comparaison. Deuxièmement, ils ont également suivi l'adoption des amis et des voisins des fermiers de comparaison. Enfin, ils ont invité les amis et les fermiers sélectionnés au hasard participant aux essais aux étapes importantes dans le développement de l'expérimentation et ont aussi suivi l'adoption ultérieure. Ces questions sont très importantes pour notre compréhension de l'adoption et de la diffusion de la technologie et la capacité à engendrer une variation exogène au travers d'une évaluation de programme par assignation aléatoire a grandement aidé à cette compréhension. De plus, leur réponse a également aidé l'ONG à développer un programme d'expansion agricole basé à l'école qui a une chance d'être efficace et d'être peu coûteux. Une version pilote de ce programme est actuellement en cours d'évaluation.

4 LE RÔLE QUE LES AGENCES INTERNATIONALES PEUVENT JOUER

La pratique courante

Les exemples discutés ci-dessus montrent qu'il est possible d'obtenir une preuve convaincante de l'impact d'un programme en organisant des projets pilotes, en profitant de l'expansion de projets existants ou en profitant de la conception du projet. Bien que tous les programmes ne puissent pas être évalués en utilisant ces méthodes, une très petite fraction de ceux qui pourraient l'être l'est. La plupart des organisations internationales exigent qu'une fraction du budget soit dépensée pour l'évaluation. Quelques pays rendent également l'évaluation obligatoire (par exemple, l'évaluation de tous les programmes sociaux est exigée par la Constitution au Mexique). Cependant, en pratique, cette part du budget n'est pas toujours dépensée efficacement : les évaluations deviennent sous-traitées à des équipes de consultants non entraînées, avec peu de conseils sur ce qu'ils doivent accomplir. Pire, elles sont parfois confiées aux organisations qui ont un intérêt dans le résultat, ainsi les évaluateurs ont une mise dans les résultats qu'ils tentent d'établir. Quand une évaluation est réellement menée, elle est généralement limitée à une évaluation de *processus* : les comptes sont vérifiés, les flux de ressources sont suivis, la livraison réelle des intrants est confirmée (par exemple, les manuels scolaires ont-ils atteint l'école ?) et des enquêtes qualitatives sont utilisées pour déterminer si les intrants ont véritablement été utilisés par les bénéficiaires (les enseignants ont-ils utilisé les manuels ?) et s'il y a une preuve *prima facie* que les bénéficiaires du programme ont été satisfaits par le programme (les enfants étaient-ils heureux ?). L'évaluation du processus est clairement essentielle et devrait aussi faire partie de n'importe quelle évaluation de programme : si aucun manuel n'a été véritablement distribué, ne trouver aucun impact du programme ne sera pas surprenant. Cependant, uniquement observer les réactions des bénéficiaires à un programme peut mener à des conclusions très trompeuses sur son efficacité : certains programmes peuvent, aux yeux de tous, sembler être des succès retentissants même s'ils n'ont pas atteint leurs objectifs. L'accent mis sur l'évaluation de processus implique que, le plus souvent, les évaluations d'impact, quand elles prennent place, viennent après coup et ne sont pas planifiées pour débuter en même temps que le programme.

Le Programme d'Education Primaire de District (DPEP), le plus gros programme d'éducation financé par la Banque Mondiale, mis

en œuvre en Inde, est un exemple d'un gros programme qui offrait un potentiel pour des évaluations très intéressantes, mais dont le potentiel de ce point de vue fut mis en péril par le manque de planification. DPEP était supposé être une vitrine de la capacité à « aller à l'échelle » avec la réforme de l'éducation (Pandey 2000). Case (2001) fournit une discussion éclairante du programme et des caractéristiques qui rendent son évaluation impossible. DPEP est un vaste programme cherchant à améliorer les performances de l'éducation publique. Il implique un entraînement pour les enseignants, des intrants et des salles de classes. Il est donné en général un haut niveau de discrétion aux districts concernant la manière de dépenser les ressources additionnelles. En dépit de l'engagement apparent en faveur d'une évaluation soigneuse du programme, plusieurs caractéristiques rendent une évaluation convaincante de l'impact du DPEP impossible. Premièrement, les districts furent sélectionnés selon deux critères : un faible *niveau* de réussite (mesuré par de faibles taux d'alphabétisation féminins) mais de forts *potentiels d'amélioration.* En particulier, les premiers districts choisis pour recevoir le programme furent sélectionnés « sur la base de leur capacité à réussir dans un laps de temps raisonnable » (Pandey 2000, cité dans Case 2001). La combinaison de ces deux éléments dans le processus de sélection rend clair que toute comparaison entre le niveau de réussite des districts DPEP et des districts non DPEP serait probablement biaisée vers le bas alors que toute comparaison entre l'amélioration de la réussite entre les districts DPEP et non DPEP (« différences dans les différences ») serait probablement biaisée vers le haut. Cela n'a pas empêché le DPEP de mettre énormément l'accent sur le suivi et l'évaluation : de grandes quantités de données ont été collectées et de nombreux rapports ont été commandés. Cependant, le processus de collecte de données ne fut mené *que dans les districts DPEP* ! Ces données peuvent uniquement être utilisées pour effectuer des comparaisons avant/après, ce qui n'a clairement aucun sens dans une économie subissant une croissance et une transformation rapides. Si jamais un chercheur trouvait une stratégie d'identification crédible, il ou elle devrait utiliser les données du recensement ou de l'Enquête sur Echantillon National.

L'économie politique de l'évaluation de programme

Nous avons affirmé que les problèmes de biais de variable omise pour le traitement desquels les évaluations par assignation aléatoire sont conçues sont réels et que les évaluations par assignation aléatoire sont faisa-

bles. Elles ne sont pas plus coûteuses que les autres types d'enquêtes et coûtent beaucoup moins cher que la poursuite de politiques inefficaces. Alors, pourquoi sont-elles aussi rares ? Cook (2001) attribue cette rareté dans l'éducation à la culture post-moderne dans les écoles d'éducation américaines qui est hostile à la conception traditionnelle de causalité sous-jacente à la mise en œuvre statistique. Pritchett (2003) affirme que les partisans d'un programme trompent systématiquement les votants indécis en les faisant croire à des estimations exagérées des impacts du programme. Les partisans bloquent les évaluations par assignation aléatoire puisqu'elles révèleraient les véritables impacts du programme aux votants. Kremer (2003) a proposé une explication complémentaire où les décideurs politiques ne sont pas systématiquement dupés mais ont des difficultés à jauger la qualité de la preuve sachant que les partisans peuvent supprimer les résultats d'une évaluation défavorable. Les partisans d'un programme sélectionnent l'estimation la plus haute pour la présenter aux décideurs politiques alors que tout opposant sélectionne l'estimation la plus négative. Sachant cela, les décideurs politiques ne tiennent rationnellement pas compte de ces estimations. Par exemple, si les partisans présentent une étude montrant un taux de rendement de 100 %, le décideur politique devrait supposer que le véritable rendement est de 10 %. Dans cet environnement, il y a peu d'incitation à mener des évaluations par assignation aléatoire : puisque les estimations qui en résultent n'incluent aucun terme de biais, elles ont peu de chances d'être suffisamment hautes ou suffisamment basses pour que les partisans les présentent aux décideurs politiques. Même si les résultats sont présentés aux décideurs politiques, ces décideurs politiques sont incapables de jauger la qualité d'études particulières et ne vont pas en tenir compte. Pourquoi financer un projet dont une évaluation par assignation aléatoire suggère qu'il a un taux de rendement de 25 % alors que les partisans de projets concurrents prétendent qu'ils ont un taux de rendement de 100 % ?

Dans ce monde, une organisation internationale peut jouer un rôle clé en encourageant les évaluations par assignation aléatoire et en les finançant. De plus, s'il devient aisé pour les décideurs politiques et les donneurs d'identifier une évaluation crédible quand il y a déjà des exemples (qui semblent plausibles), ce rôle peut en fait débuter un cercle vertueux en encourageant les autres donneurs à reconnaître et avoir confiance en une évaluation crédible et alors plaider pour générer une telle évaluation par opposition à d'autres. De la sorte, elles peuvent contribuer à un « climat » favorable à une évaluation crédible et ainsi surmonter la réticence que nous avons mentionnée plus haut. Le procédé d'évaluation qualitative lui-même serait alors augmenté au-dessus et

au-delà de ce que les organisations peuvent elles-mêmes promouvoir et financer.

L'évaluation dans les organisations internationales

Les organisations internationales peuvent jouer plusieurs rôles dans la promotion et le financement d'évaluations rigoureuses.

Il est presque certainement contreproductif d'exiger que *tous les projets* soient sujets à des évaluations d'impact. Clairement, tous les projets ont besoin d'être surveillés pour être sûr qu'ils ont réellement lieu et pour éviter un mauvais emploi des fonds. Cependant, certains programmes ne peuvent simplement pas être évalués avec les méthodes discutées dans cet article. Et même parmi les projets qui pourraient potentiellement être évalués, tous n'ont pas besoin d'évaluations d'impact. En fait, la valeur d'une évaluation d'impact pauvrement identifiée est très faible et son coût, en termes de crédibilité, est élevé, spécialement si les organisations internationales prennent un rôle moteur dans la promotion d'une évaluation de qualité. Un premier objectif est donc de réduire le nombre d'évaluations ruineuses; toute évaluation d'impact proposée devrait être revue par un comité avant qu'une quelconque somme ne soit dépensée en collecte de données. La responsabilité du comité serait d'établir la capacité de l'évaluation à livrer des estimations causales fiables de l'impact du projet. Un second objectif serait de mener des évaluations crédibles dans des domaines clés. En consultation avec un corps de chercheurs et de praticiens, chaque organisation devrait déterminer les domaines clés dans lesquels elle promouvra des évaluations d'impact. Des évaluations par assignation aléatoire pourraient aussi être mises en place dans d'autres domaines quand l'opportunité se présente.

Des évaluations d'impact crédibles requièrent beaucoup de travail et, en plus, les bénéfices d'évaluations d'impact crédibles s'étendent bien au-delà de l'organisation menant l'évaluation; ces facteurs signifient que les incitations à mener des évaluations rigoureuses sont moins qu'optimales socialement. Un remède prometteur est d'insérer au sein du cadre institutionnel des agences internationales des structures qui fourniront suffisamment d'incitations pour les évaluateurs. Étant donné l'actuelle rareté des évaluations par assignation aléatoire au sein de l'environnement institutionnel des organisations internationales, il peut y avoir de la place pour mettre en place un fonds spécialisé pour encourager et financer des évaluations d'impact rigoureuses et pour diffuser les résultats. Comme nous en discuterons brièvement plus loin, le potentiel pour un tel fonds est énorme : il existe une offre potentielle

d'évaluateurs toute faite à la fois au sein des agences internationales elles-mêmes et au sein du milieu universitaire et les collaborations avec les ONG offrent de nombreuses opportunités pour évaluer des politiques de grande pertinence.

Un tel fonds d'évaluation encouragerait la collecte de données et l'étude de véritables « évaluations par assignation aléatoire naturelles » avec une assignation aléatoire induite par le programme. Comme nous l'avons mentionné dans la section 2 ci-dessus, les évaluations par assignation aléatoire ne sont pas la seule méthode pour mener de bonnes évaluations d'impact. Cependant, de telles autres évaluations sont menées beaucoup plus couramment alors que les évaluations par assignation aléatoire sont menées beaucoup trop rarement au regard de leur valeur et des opportunités de les mener. Une partie du problème vient du fait que personne ne considère la réalisation de telles évaluations comme étant son travail et donc personne n'investit suffisamment pour les mener. En outre, toutes les évaluations ont des caractéristiques communes et bénéficieraient donc d'une unité spécialisée avec une expertise spécifique. Puisque les évaluations d'impact engendrent des biens publics internationaux, l'unité devrait avoir un budget qui serait utilisé pour financer et mener des évaluations rigoureuses de projets internes et externes. L'unité devrait mener ses propres projets d'évaluation dans les domaines clés identifiés par l'organisation.

Comme cela a été discuté précédemment, l'unité devrait également travailler en partenariat, particulièrement avec les ONG et les universitaires. Pour les projets soumis de manière extérieure à l'unité, un comité au sein de l'unité (potentiellement avec l'aide de critiques extérieurs) pourrait recevoir des propositions émanant de l'organisation ou de personnes extérieures et à partir de là choisir les projets à supporter. L'unité pourrait également encourager la réplique d'évaluations importantes en envoyant des appels pour des propositions spécifiques. Le projet pourrait alors être mené en partenariat avec des personnes issues de l'unité ou d'autres chercheurs (en particulier des universitaires). L'unité pourrait fournir à la fois un support financier et technique pour le projet avec du personnel et des chercheurs attitrés. Avec le temps, sur la base de l'expérience acquise, l'unité pourrait également servir de centre de ressource plus général en développant et en diffusant des modules d'entraînement, des outils et des directives (des instruments d'enquêtes et de test de même qu'un logiciel qui peut être utilisé pour la saisie de données et pour faciliter l'assignation aléatoire – similaires dans l'esprit aux outils produits par d'autres unités de la Banque Mondiale) pour l'évaluation par assignation aléatoire. L'unité pourrait également

parrainer des sessions d'entraînement pour les praticiens. Un autre rôle que l'unité pourrait jouer, après avoir établi une réputation de qualité, est celui d'agence de diffusion (une « chambre de compensation » en quelque sorte). Pour être utiles, les résultats d'évaluation doivent être accessibles aux praticiens à la fois au sein et hors des agences de développement. Un rôle clé de l'unité pourrait être de mener des recherches systématiques pour toutes les évaluations d'impact, établir leur fiabilité et publier les résultats sous la forme de dossiers de politique dans une base de données consultables facilement accessible. Dans l'idéal, la base de données inclurait toute information qui pourrait être utile à l'interprétation des résultats (estimations, taille de l'échantillon, région et époque, type de projet, coût, analyse coût-bénéfice, avertissements, etc.) ainsi que les références aux études en rapport. La base de données pourrait inclure à la fois les évaluations d'impact avec et sans assignation aléatoire satisfaisant certains critères à condition que les différents types d'évaluation soient clairement libellés. Les évaluations devraient satisfaire des exigences de rapport minimum pour être incluses dans la base de données et tous les projets supportés par l'unité devraient être inclus dans la base de données, quels que soient leurs résultats.

Comme cela a été discuté précédemment, une telle base de données aiderait à réduire le biais de publication, qui peut être substantiel si les résultats positifs ont plus de chances d'être publiés. Les journaux académiques ne sont sans doute pas intéressés par la publication des résultats d'un programme ayant échoué mais du point de vue des décideurs politiques, la connaissance de résultats négatifs est tout simplement aussi utile que la connaissance de projets ayant réussi. Des exigences comparables sont posées sur tous les projets médicaux financés au niveau fédéral aux États-Unis. Dans l'idéal, avec le temps, la base de données devrait devenir une référence de base pour les organisations et les gouvernements, particulièrement au moment où ils cherchent un financement pour leurs projets. Cette base de données pourrait lancer un cercle vertueux, avec les donneurs exigeant des évaluations crédibles avant de financer ou de continuer des projets, plus d'évaluations étant menées et la qualité générale du travail d'évaluation augmentant.

5 CONCLUSION : UTILISER L'ÉVALUATION POUR CONSTRUIRE UN CONSENSUS DE LONG TERME POUR LE DÉVELOPPEMENT

Des évaluations rigoureuses et systématiques ont le potentiel pour pousser l'impact des organisations internationales bien au-delà de leur simple capacité à financer des programmes. Des évaluations d'impact crédibles sont des biens publics internationaux : les bénéfices du fait de savoir si un programme fonctionne ou ne fonctionne pas s'étendent au-delà de l'organisation ou du pays mettant en œuvre le programme. Les programmes qui ont montré qu'ils avaient réussi peuvent être adaptés pour être utilisés dans d'autres pays et élargis au sein des pays, alors que les programmes ayant échoué peuvent être abandonnés. Au travers de la promotion, de l'encouragement et du financement d'évaluations rigoureuses (telles que des évaluations par assignation aléatoire crédibles) des programmes qu'elles supportent aussi bien que des programmes supportés par les autres, les organisations internationales peuvent fournir des conseils aux organisations internationales elles-mêmes aussi bien qu'aux autres donneurs, aux gouvernements et aux ONG dans la recherche en cours pour les programmes qui réussissent et ainsi améliorer l'efficacité de l'aide au développement. De plus, en établissant de manière crédible les programmes qui fonctionnent et ceux qui ne fonctionnent pas, les agences internationales peuvent contrebalancer le scepticisme concernant la possibilité de dépenser l'aide efficacement et construire un support de long terme pour le développement. Il s'agit de l'opportunité de réaliser un véritable « changement d'échelle ».

RÉFÉRENCES

ANGRIST, Joshua, Eric BETTINGER, Erik BLOOM, Elizabeth KING et Michael KREMER (2002), "Vouchers for Private Schooling in Colombia : Evidence from a Randomized Natural Experiment", *American Economic Review*, 92(5), pp. 1535-58.

ANGRIST, Joshua et Alan KRUEGER (1999), "Empirical strategies in labor economics", *in* Orley Ashenfelter et David Card, eds., *Handbook of Labor Economics, Vol. 3A*, Amsterdam, North Holland, pp. 1277-1366.

ANGRIST, Joshua et Alan KRUEGER (2001), "Instrumental Variables and the Search for Identification : From Supply and Demand to Natural Experiments" *Journal of Economic Perspectives* 15 (4), pp. 69-85.

ANGRIST, Joshua, et Victor LAVY (1999), "Using Maimonides' rule to estimate the effect of class size on scholastic achievement" *Quarterly Journal of Economics*, 114 (2), pp. 533-575.

ASHENFELTER, Orley, Colm HARMON et Hessel OOSTERBEEK (1999), "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias", *Labour Economics*, 6 (4), pp. 453-70.

ATTANASIO, Orazio, Costas MEGHIR, et Ana SANTIAGO (2001), "Education Choices in Mexico : Using a Structural Model and a Randomized Experiment to evaluate PROGRESA", Mimeo, Inter-American Development Bank.

BANERJEE, Abhijit (2002), "The Uses of Economic Theory : Against a Purely Positive Interpretation of Theoretical Results", Mimeo, MIT.

BANERJEE, Abhijit, Shawn COLE, Esther DUFLO et Leigh LINDEN (2003), "Remedying Education : Evidence from Two Randomized Experiments", Mimeo, MIT.

BANERJEE, Abhijit et Ruimin HE (2003), "The World Bank of the Future", *American Economic Review, Papers and Proceedings*, 93 (2), pp. 39-44.

BANERJEE, Abhijit, Suraj JACOB, et Michael KREMER (avec Jenny Lanjouw et Peter Lanjouw) (2001), "Promoting School Participation in Rural Rajasthan : Results from Some Prospective Trials", Mimeo, Harvard-MIT.

BEHRMAN, Jere, Piyali SENGUPTA et Petra TODD (2002), "Progressing through PROGESA : An Impact Assessment of a School Subsidy Experiment in Mexico", Mimeo, University of Pennsylvania.

BERTRAND, Marianne, Esther DUFLO et Sendhil MULLAINATHAN (2003), "How Much Should We Trust Difference in Differences Estimates ?", à paraître *in Quarterly Journal of Economics.*

BESLEY, Timothy et Anne CASE (2000), "Unnatural Experiments ? Estimating the Incidence of Endogenous Policies", *Economic Journal*, 110 (467), pp. F672-F694.

BOBONIS, Gustavo, Edward MIGUEL, et Charu SHARMA (2002), "Iron Supplementation and Early Childhood Development : A Randomized Evaluation in India", Mimeo, University of California, Berkeley.

BUDDLEMEYER, Hielke et Emmanuel SKOFIAS (2003), "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA", Institute for Study of Labor, Discussion Paper No. 827.

CAMPBELL, Donald T. (1969), "Reforms as Experiments", *American Psychologist*, 24, pp. 407-429.

CARD, David (1999), "The causal effect of education on earnings", *in* Orley Ashenfelter et David Card, eds., *Handbook of Labor Economics, Vol. 3A*, Amsterdam, North Holland, pp. 1801-63.

CASE, Anne (2001), "The primacy of education." Mimeo, Princeton University.

CHATTOPADHYAY, Raghabendra et Esther DUFLO (2001), "Women as Policy Makers : Evidence from a India-Wide Randomized Policy Experiment." NBER Working Paper # 8615.

CHATTOPADHYAY, Raghabendra et Esther DUFLO (2003), "Women as Policy Makers : Evidence from a India-Wide Randomized Policy Experiment", Mimeo, MIT.

COOK, Thomas D. (2001), "Reappraising the Arguments Against Randomized Experiments in Education : An Analysis of the Culture of Evaluation in American Schools of Education", Mimeo, Northwestern University.

CRONBACH, L. (1982), *Designing evaluations of educational and social programs*, San Francisco, Jossey-Bass.

CRONBACH, L., S. AMBRON, S. DORNBUSCH, R. HESS, R. HORNIK, C. PHILLIPS, D. WALKER et S. WEINER (1980), *Toward reform of program evaluation*, San Francisco, Jossey-Bass.

CULLEN, Julie Berry, Brian JACOB et Steven LEVITT (2002), "Does School Choice Attract Students to Urban Public Schools ? Evidence from over 1,000 Randomized Lotteries", Mimeo, University of Michigan.

DELONG, J. BRADFORD, et Kevin LANG (1992), "Are All Economic Hypotheses False ?", *Journal of Political Economy*, 100(6), pp. 1257-72.

DUFLO, Esther (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia : Evidence from an Unusual Policy Experiment", *American Economic Review*, 91(4), pp. 795-813.

DUFLO, Esther (2003), "Poor but Rational ?", Mimeo, MIT.

DUFLO, Esther, Pascaline DUPAS, Michael KREMER et Samuel SINEI (2003), "Evaluating HIV/AIDS prevention education in primary schools : Preliminary results from a randomized controlled trial in Western Kenya", Mimeo, Harvard-MIT.

DUFLO, Esther et Michael KREMER (2003), "Use of Randomization in the Evaluation of Development Effectiveness", Mimeo, MIT.

GALASSO, Emanuela, Martin RAVALLION, et Agustin SALVIA (2002), "Assisting the Transition from Workfare to Work : A Randomized Experiment", Mimeo, Development Research Group, World Bank.

GERTLER, Paul J., et Simone BOYCE (2001), "An experiment in incentive-based welfare : The impact of PROGRESA on health in Mexico", Mimeo, University of California, Berkeley.

GLEWWE, Paul, Nauman ILIAS, et Michael KREMER (2003), "Teacher Incentives", Mimeo, Harvard University.

GLEWWE, Paul, Michael KREMER, Sylvie MOULIN et Eric ZITZEWITZ (2003), "Retrospective vs. prospective analyses of school inputs : The case of flip charts in Kenya", à paraître dans *Journal of Development Economics.*

HECKMAN, James, Lance LOCHNER, et Christopher TABER (1998), "General Equilibrium Treatment Effects : A Study of Tuition Policy", NBER Working Paper #6426.

HSIEH, Chang-Tai et Miguel URQUIOLA (2002), "When Schools Compete, How Do They Compete ? An assessment of Chile's nationwide school voucher program", Mimeo, Princeton University.

IMBENS, Guido, et Joshua ANGRIST (1994), "Identification and estimation of local average treatment effects", *Econometrica*, 62(2), pp. 467-475.

KREMER, Michael (2003), "Randomized Evaluations of Educational Programs in Developing Countries : Some Lessons", *American Economic Review, Papers and Proceedings*, 93(2), pp. 102-115.

KRUEGER, Alan (1999), "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics* 114(2), pp. 497-532.

LALONDE, Robert (1986), "Evaluating the Econometric Evaluations of Training with Experimental Data", *American Economic Review*, 76(4), pp. 604-620.

MEYER, Bruce D. (1995), "Natural and quasi-experiments in economics", *Journal of Business and Economic Statistics*, 13(2), pp. 151-161.

MIGUEL, Edward, et Michael KREMER (2003), "Worms : Identifying Impacts on Education and Health in the Presence of Treatment Externalities", à paraître *in Econometrica.*

MORDUCH, Jonathan (1998), "Does microfinance really help the poor ? New evidence from flagship programs in Bangladesh", Mimeo, Princeton University.

NARAYANAN, Deepa, ed. (2000), *Empowerment and Poverty Reduction : A Sourcebook*, Washington DC, The World Bank.

PANDEY, Raghaw Sharan (2000), *Going to Scale With Education Reform : India's District Primary Education Program, 1995-99. Education Reform and Management Publication Series, Volume I, No. 4*, Washington DC, World Bank.

PITT, Mark et Shahidur KHANDKER (1998), "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh : Does the Gender of Participants Matter ?", *Journal of Political Economy*, 106(5), pp. 958-996.

PRITCHETT, Lant (2003), "It Pays to be Ignorant : A Simple Political Economy of Rigorous Program Evaluation", à paraître, *Journal of Policy Reform.*

ROSENBAUM, Paul R. (1995), "Observational studies", dans *Series in Statistics*, New York, Heidelberg, London, Springer.

SHULTZ, T. Paul (2001), "School Subsidies for the Poor : Evaluating the Mexican PROGRESA Poverty Program", à paraître, *Journal of Development Economics.*