International Initiative for Impact Evaluation

WORKING PAPER 3
IN FRENCH

L'évaluation d'impact basée sur la théorie : principes et pratique

Howard White
Juin 2009

À propos de 3ie

L'Initiative internationale pour l'évaluation d'impact (3ie) œuvre à l'amélioration de la vie des populations du monde en développement en soutenant la production et l'utilisation de données probantes sur ce qui marche, quand, pourquoi et à quel prix. 3ie est une nouvelle initiative qui répond à la demande de données de plus grande qualité et qui renforcera l'efficacité du développement en encourageant des politiques mieux informées. 3ie finance des évaluations d'impact de qualité et milite pour éclairer la conception des programmes et des politiques dans les pays en développement.

La série Documents de travail de 3ie couvre à la fois des questions conceptuelles relatives à l'évaluation d'impact et des constats d'études spécifiques ou de revues de synthèse.

Ce document de travail a été écrit par Howard White, Directeur exécutif de 3ie.

Contacts

Initiative internationale pour l'évaluation d'impact
c/o Global Development Network
Post Box No. 7510
Vasant Kunj P.O.
New Delhi – 110070, Inde
Tél. : +91-11-2613-9494/6885
www.3ieimpact.org

L’évaluation d’impact basée sur la théorie : principes et pratique

Howard White
Directeur exécutif
Initiative internationale pour l’évaluation d’impact, 3ie
Contact : hwhite@3ieimpact.org

Résumé

La demande d’évaluations d’impact rigoureuses a conduit à rechercher comment déterminer les interventions fructueuses, mais aussi pourquoi elles le sont. On s’accorde aujourd’hui à penser que la question du pourquoi est éclairée par une approche de l’évaluation d’impact basée sur la théorie, qui cartographie la chaîne causale des moyens aux résultats et à l’impact et vérifie les hypothèses sous-jacentes. Pourtant, cette approche reste mal appliquée. Cet article recense six principes pour une application fructueuse : (1) cartographie de la chaîne causale (théorie du programme) ; (2) compréhension du contexte ; (3) anticipation de l’hétérogénéité ; (4) évaluation rigoureuse de l’impact au moyen d’un contrefactuel crédible ; (5) analyse factuelle rigoureuse ; (6) mixité des méthodes.

1. Introduction

Le recours aux méthodes quantitatives pour mesurer l'impact des programmes de développement suscite depuis quelques années un intérêt accru. Grâce au programme de travail d'organisations telles que le Poverty Action Lab (J-PAL) et Innovations in Poverty Action (IPA),[1] au portefeuille d'études financées dans le cadre de l'Initiative pour l'évaluation de l'impact au plan du développement (DIME) et du Fonds espagnol d'évaluation d'impact (SIEF) de la Banque mondiale[2], mais aussi aux financements consentis par l'Initiative internationale pour l'évaluation d'impact (3ie)[3], on disposera d'ici à cinq ans de centaines d'études de ce type alors que les revues entreprises ces dernières années n'en mentionnaient que quelques-unes (voir par exemple, Centre for Global Development, 2006). Cependant, pour la plupart des partisans de l'amélioration des évaluations d'impact, savoir ce qui marche n'est pas suffisant, il faut comprendre pourquoi. Or il ne suffit pas pour cela de rapporter l'effet moyen du traitement d'une intervention, d'où la déclaration du Réseau de réseaux pour l'évaluation d'impact (NONIE) : « l'application de l'approche basée sur la théorie implique qu'une évaluation bien conçue couvre à la fois les questions d'évaluation des processus et de l'impact. La pertinence pour les politiques publiques s'en trouve renforcée car l'étude permet de déterminer si une intervention a eu l'impact recherché, mais aussi pourquoi elle l'a eu – ou non » (NONIE, non daté). De même, le guide de 3ie sur la pratique de l'évaluation d'impact déclare que «les études doivent clairement exposer comment l'intervention (les moyens) est censée affecter les résultats finaux et vérifier chaque lien (hypothèse) entre les moyens et les résultats (ce qu'on appelle parfois la théorie du programme). Le protocole d'évaluation doit prévoir l'analyse de la chaîne causale des moyens aux impacts » (3ie, non daté).

L'approche préconisée ici pour comprendre pourquoi un programme a ou n'a pas eu un impact est appelée « évaluation d'impact basée sur la théorie ». L'idée n'est pas nouvelle. Cette approche, qui suppose d'examiner les hypothèses qui sous-tendent la chaîne causale des moyens aux résultats et à l'impact, est en effet déjà ancienne (voir par exemple, Weiss 1998, et Carvalho et White, 2004, pour une application au développement). Certains praticiens des essais contrôlés randomisés (ECR) et des méthodes quasi expérimentales recourent depuis longtemps à la théorie du programme pour expliquer leurs constats (Blackman et Reich, 2009, p. 67-68). D'autre part, dans son article examinant les protocoles d'évaluation d'impact possibles pour un ensemble d'interventions de développement, Rogers (2009) note qu'une approche basée sur la théorie serait appropriée à tous les cas.

Malgré un engagement de principe en faveur de l'évaluation basée sur la théorie, peu d'études semblent tenir les promesses de cette approche. Cet article se propose de contribuer à remédier à ces insuffisances en exposant les étapes, ou principes, de l'évaluation basée sur la théorie. Cette introduction est suivie d'une 2e partie, qui présente un exemple, le Projet de nutrition intégrée du Bangladesh (*Bangladesh Integrated Nutrition*

[1] Voir www.povertyactionlab.org et http://poverty-action.org respectivement.
[2] Voir www.worldbank.org/dime et www.worldbank.org/sief respectivement.
[3] Voir www.3ieimpact.org.

Project, BINP), dans lequel je puise ensuite, avec d'autres exemples, pour illustrer les principes analysés dans la 3e partie. L'article se poursuit par une brève comparaison entre l'évaluation d'impact basée sur la théorie et les approches de type boîte noire dans la 4e partie, et conclut à la 5e partie.

2. Exemple – le Projet de nutrition intégrée du Bangladesh

Cette partie présente brièvement l'évaluation du Projet de nutrition intégrée du Bangladesh (*Bangladesh Integrated Nutrition Project*, BINP), un cas qui servira plus loin pour illustrer les principes qui sous-tendent l'évaluation d'impact basée sur la théorie, qui sont analysés dans la partie suivante. Pour une étude plus approfondie de ce projet, le lecteur pourra se reporter à Banque mondiale (2005), White et Masset (2006) et White (2005).

Le BINP, qui s'est inspiré du Projet de nutrition intégrée du Tamil Nadu (*Tamil Nadu Integrated Nutrition Project*, TINP) en Inde, dont le succès a été salué, était un projet de suivi de la croissance des enfants, dans le cadre duquel les jeunes enfants étaient pesés toutes les semaines dans un centre de pesée locale tenu par une villageoise formée pour être agent communautaire de nutrition. Celle-ci établissait une courbe de croissance par rapport à l'âge à partir des relevés de poids. Les enfants qui ne grandissaient pas assez (retards de croissance) ou qui prenaient trop de retard par rapport à la norme de référence (malnutris), étaient recrutés au sein du programme, lequel consistait en conseils nutritionnels et en distribution de suppléments alimentaires. Cependant, la documentation du projet indiquait clairement que les séances de conseil devaient constituer le premier facteur d'impact. L'idée était que les problèmes de nutrition tenaient davantage à l'ignorance qu'à la pauvreté, argument étayé par des données révélant un problème de malnutrition même dans le quintile le plus riche, et à des croyances voulant par exemple que les femmes enceintes mangent moins pendant leur grossesse. Le programme devait également dispenser des conseils en nutrition et des suppléments alimentaires aux femmes enceintes. Le BINP, qui était un programme pilote, a été remplacé par la suite par le Programme de nutrition national (*National Nutrition Program,* NNP).

On a d'abord pensé que le BINP était une réussite car les données de suivi montraient un important recul de la malnutrition, en particulier dans ses formes graves, dans les régions du projet. Sur la base de ces données, la Banque mondiale a décidé de déployer le projet à l'échelle nationale, le NNP, à peu près à mi-parcours du BINP et avant toute évaluation. Cette décision a été critiquée dans un rapport publié par l'organisation Save the Children UK, dont les données obtenues par simple comparaison ex post des régions de traitement et de contrôle ne constatait aucune différence entre les deux (Save the Children, 2003).

L'analyse entreprise par le Département de l'évaluation des opérations de la Banque mondiale (OED, remplacé par le Groupe d'évaluation indépendante, IEG), utilisait l'appariement des scores de propension, qui combinait des données des régions du projet avec des données issues d'une enquête nationale sur la nutrition réalisée par Helen Keller International pour constituer un groupe de contrôle. Cette analyse n'a pas constaté d'impact

significatif sur l'état nutritionnel, bien qu'il y ait eu un impact positif sur les enfants les plus malnutris.

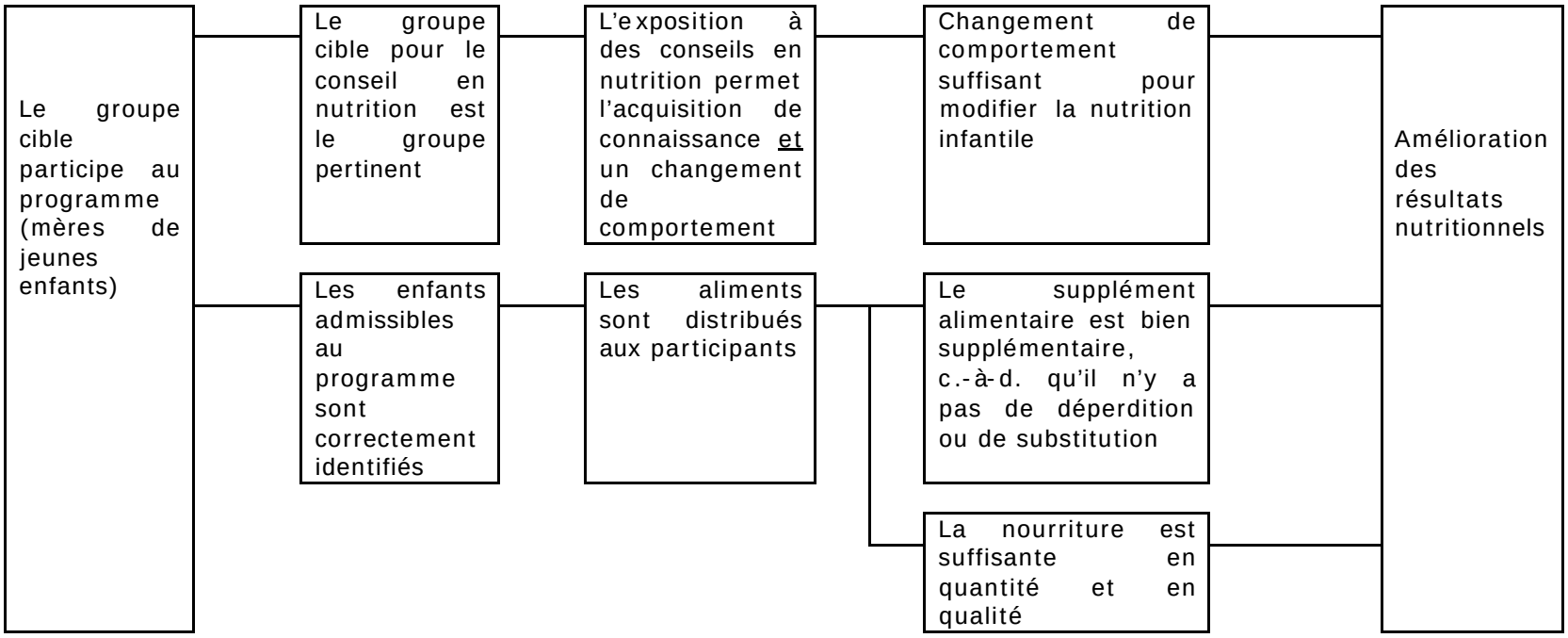
Un certain nombre d'hypothèses intégrées à la chaîne causale ont pu conduire à anticiper un impact positif du BINP sur les résultats nutritionnels, dont certaines sont présentées à la Figure 1.

La première est que les individus ont effectivement connaissance du programme et y participent – de nombreux projets de développement échouent au premier obstacle en raison d'efforts insuffisants pour expliquer l'intervention aux bénéficiaires ciblés ou pour effectuer une évaluation réaliste des coûts et avantages relatifs pour les bénéficiaires. En fait, le BINP a eu de bons résultats sur ce plan, près de 90 % des femmes admissibles ayant amené leurs enfants, même si comme nous le verrons plus loin, d'importantes exceptions sont à noter.

Deuxièmement, les individus ciblés sont les bons. Le programme ciblait les mères de jeunes enfants. Or bien souvent, ce ne sont pas les mères qui décident, et en tout état de cause, elles sont rarement les seuls décideurs en matière de santé et de nutrition des enfants. D'une part, ce ne sont pas les femmes qui vont au marché dans le Bangladesh rural, mais les hommes. D'autre part, pour les femmes qui vivent au sein d'une famille étendue – c'est-à-dire avec leur belle-mère – comme le font une importante minorité d'entre elles, c'est la belle-mère qui dirige le domaine des femmes. De fait, les taux de participation au projet sont nettement plus bas pour les femmes qui vivent avec leur belle-mère dans les régions les plus traditionnelles du pays.

Une fois que les femmes se présentent pour faire peser leurs enfants, il faut recruter les bons enfants, c'est-à-dire ceux qui ont des troubles de croissance ou qui sont malnutris. Mais les données ont montré d'importantes erreurs de ciblage, tant de type I (enfants non recrutés alors qu'ils devraient l'être) que de type II (enfants recrutés qui n'auraient pas dû l'être). Nous avons testé les agents communautaires de nutrition avec quelques exemples de courbes de croissance (celles qui étaient utilisées pour la formation) et il s'est avéré que la plupart ne savaient pas déterminer quels enfants il fallait recruter à partir des courbes, d'où les erreurs de ciblage. C'est très important pour l'impact du programme car nous avons constaté que les enfants les plus malnutris en bénéficiaient effectivement, de sorte que l'impact moyen aurait été plus sensible si le programme s'était concentré sur ces enfants, alors qu'en fait les ressources étaient affectées à des enfants qui n'en bénéficieraient pas.

Figure 1 – Chaîne causale du projet de nutrition : conseil en nutrition et supplémentation alimentaire



En outre, pour que l'alimentation supplémentaire ait un impact positif, elle doit être supplémentaire, alors qu'en réalité il y avait à la fois des déperditions (les aliments étaient donnés à une autre personne que leur destinataire, en particulier les suppléments donnés aux femmes enceintes) et des substitutions (les aliments étaient pris en remplacement d'un repas qui aurait été donné en l'absence du programme).

Quant au changement de comportement, il y a bien eu communication sur le changement comportemental, mais les comportements n'ont pas changé ; les femmes inscrites au programme étaient nettement mieux informées sur les « bonnes pratiques », mais il y avait un important écart entre la théorie et la pratique : de nombreuses femmes n'appliquaient pas ces connaissances. Cette situation tenait en partie à des contraintes de ressources : les femmes des ménages pauvres étaient moins susceptibles de s'alimenter davantage pendant la grossesse et celles qui vivaient dans des ménages possédant du terrain ou avec un parent mâle plus âgé étaient moins susceptibles de prendre davantage de repos pendant leur grossesse. Mais l'autre raison était encore une fois les belles-mères. Un focus group a explicitement déclaré aux investigateurs que « lorsque nos belles-mères seront mortes, alors peut-être que nous ferons ce que vous dites, mais d'ici là, nous suivrons les traditions ». Enfin, la probabilité que certains changements de comportement – en particulier ceux visant à accroître la prise de poids pendant la grossesse – aient un impact sensible sur le résultat final du poids de naissance était faible (c'est le poids de la mère avant la grossesse qui est le plus important pour cela).

Bref, l'impact du projet était sapé par l'absence de liens ou la faiblesse des liens dans la chaîne causale. Globalement, le projet n'avait pas d'impact. Les améliorations révélées par les données de suivi du projet se produisaient en fait dans tout le pays. Save the Children n'a pas constaté de différence entre les régions dans lesquelles le projet avait été déployé et les régions de contrôle. En fait, les améliorations tendancielles observées trouvaient leur origine dans l'augmentation des rendements du riz, une élévation des revenus et la baisse des prix du riz, et pas dans le BINP.

Cependant, l'analyse a pointé des pistes très claires d'amélioration des performances du programme : (1) participation des belles-mères et des maris aux séances de conseils en nutrition, (2) ciblage plus étroit du programme et (3) amélioration du ciblage par une meilleure formation des agents communautaires de nutrition, et peut-être un recrutement plus sélectif des agents. Cependant, les calculs ont également montré qu'il s'agissait d'une intervention très coûteuse – qu'il serait difficile de déployer à grande échelle du fait de contraintes de gestion et de ressources.

Malheureusement, les leçons tirées de cette évaluation n'ont pas été suivies d'effets. Très attachée au modèle du TINP/BINP, l'équipe de la Banque mondiale chargée de la nutrition a considéré que le TINP était un franc succès (alors qu'aucune étude rigoureuse n'a été conduite selon les normes actuelles) et la Banque mondiale a également revendiqué une réussite au Bangladesh, alors que cette affirmation était contestée. Après

avoir pris part à ces débats, l'équipe nutrition de la Banque mondiale a publié un document qui concluait sans réserves au succès du BINP (Banque mondiale, 2006). Partant de cette conviction, il fut décidé de déployer le NNP en appliquant le modèle du BINP, alors que les données d'évaluation montraient que le modèle qui avait peut-être fonctionné à Tamil Nadu nécessitait des adaptations pour fonctionner au Bangladesh. Trois ans plus tard, le NNP tournait court et l'étude d'impact envisagée était abandonnée car l'absence d'impact était manifeste du fait des carences de la mise en œuvre. Le rapport d'achèvement de la Banque mondiale concluait par des recommandations de réforme du programme remarquablement semblables à celles qu'avait formulées le rapport de l'OED deux ans plus tôt.

3. Principes

Les six principes fondamentaux d'une évaluation d'impact basée sur la théorie sont les suivants :

1. cartographie de la chaîne causale (théorie du programme)
2. compréhension du contexte
3. anticipation de l'hétérogénéité
4. évaluation rigoureuse de l'impact au moyen d'un contrefactuel crédible
5. analyse factuelle rigoureuse
6. mixité des méthodes employées

Cartographie de la chaîne causale (théorie du programme)

La chaîne causale lie les moyens aux résultats et aux impacts ; elle représente la théorie du programme (ou théorie du changement) relative aux mécanismes par lesquels l'intervention est censée avoir l'impact recherché. Cette théorie est intégrée au cadre logique, bien que ce dernier n'explicite pas toujours les hypothèses sous-jacentes, alors que la vérification des hypothèses est fondamentale dans une approche basée sur la théorie.

On reproche souvent à la chaîne causale d'être linéaire, c'est-à-dire soit d'être unidirectionnelle, soit d'être déterministe ; voir White (2009) pour une analyse des différents sens de « linéaire » dans le discours consacré à l'évaluation. Mais aucune de ces critiques n'est fondée. S'il peut être vrai que les responsables des programmes envisagent souvent un cadre assez simple liant les moyens aux activités aux produits, aux résultats et aux impacts, l'évaluation basée sur la théorie vérifie les hypothèses sous-jacentes à cette chaîne de raisonnement. L'un de ces liens supposés est que les résultats observés résultent des activités et des produits des projets, et non l'inverse. Mais cette causalité inverse, ou bidirectionnelle, est au cœur des débats sur l'évaluation d'impact : le biais de sélection dû au recrutement opéré par le programme et à l'autosélection des participants implique précisément que les variables de résultat affectent l'identité des participants, et non l'inverse. Ainsi, les communautés dotées d'un

niveau élevé de capital social sont plus enclines à solliciter des fonds pour des programmes de développement communautaire. Ces programmes sont destinés à renforcer le capital social, mais l'observation d'une différence ex post dans le niveau de capital social entre les villages de traitement et les villages de contrôle a plus de chances d'indiquer des différences antérieures au programme que l'impact effectif de ce dernier.

Une critique plus valide est que cette approche peut être assez statique alors qu'en règle générale, les interventions s'adaptent et évoluent. Il arrive que les systèmes décrits dans la documentation du projet aient très peu de liens avec sa mise en œuvre concrète, soit que le projet a été repensé, soit que les responsables de terrain ont interprété assez librement ses procédures. Dans le premier cas, la théorie du programme devrait refléter le nouveau protocole et le document d'évaluation le processus d'apprentissage qui a conduit à ce protocole. Dans le second cas, tout écart entre ce qui doit être fait en principe et ce qui est fait en réalité est une question fondamentale de l'évaluation : pourquoi ces différences sont-elles apparues et quels sont leurs effets sur les performances du programme ?

L'étude des fonds sociaux mentionnée plus haut donne un exemple de l'apprentissage en cours de projet. Une autre critique des investissements financés par des fonds sociaux est qu'ils ne sont pas viables car ils ne prévoient pas le fonctionnement et l'entretien. À l'origine, les fonds sociaux utilisaient un comité central qui approuvait toutes les demandes, partant du principe que les ministères de tutelle, par leur présence au comité, s'engageaient à couvrir les dépenses de fonctionnement lorsqu'ils validaient un projet. Mais ce système n'a pas marché, si bien que les fonds sociaux ont commencé à conclure des « accords-cadres » couvrant tous les projets pour chaque ministère de tutelle. Ce système présentait lui aussi des insuffisances, si bien que certains fonds sociaux ont sollicité l'accord du ministère de tutelle au cas par cas, d'autres ont exigé des plans de viabilité locaux, tandis que d'autres encore ont mis de côté des ressources pour un fonds d'entretien (voir Banque mondiale, 2002).

La théorie du programme doit être dynamique pour permettre de tirer les leçons de l'expérience sur le terrain, ce qui est une autre façon de dire qu'une itération est nécessaire entre la théorie et les données. Les méthodes d'analyse statistique basées sur un modèle prennent le modèle pour acquis et se contentent de vérifier dans quelle mesure les données sont appropriées à celui-ci – et les praticiens ont divers moyens de s'assurer que les données sont bien appropriées, puisque « les données avoueront pour peu qu'on les torture suffisamment longtemps » (Coase, cité par Leamer, 1983). Cependant, une approche par l'analyse des données permet aux données de conduire la théorie, en recherchant des configurations dans les données. Cette méthode paraît manquer de structure mais bien entendu aucun exercice statistique ne peut être dénué de théorie, puisque c'est d'abord elle qui guide la sélection des données qui seront recueillies et analysées. En réalité, la théorie doit être prête à s'adapter aux surprises que réservent les données. Cette approche peut sembler s'apparenter à l'exploration de données *(data mining)*, mais elle est en fait très différente. Dans l'exploration de données, on sait ce qu'on recherche et on poursuit l'exploration jusqu'à ce qu'on trouve.

Dans l'analyse des données en revanche, on recherche l'émergence de configurations, attendues et inattendues, dans les données (voir Mukherjee *et al.*, 1998, chapitre 1 pour une étude plus complète).

Une autre critique, peut-être fondée, est qu'en se concentrant sur la chaîne causale, l'étude passera à côté des effets non recherchés. Deux moyens permettent d'éviter cet écueil. Tout d'abord, une application soigneuse de la théorie du programme peut déceler les conséquences involontaires éventuelles ; on pourra par exemple envisager en détail les implications environnementales du programme, sur lesquelles les concepteurs ne se seront peut-être pas arrêtés. Ensuite, un travail de terrain préliminaire, comprenant une analyse participative, est une partie importante du protocole d'évaluation qui peut discerner ces résultats involontaires, lesquels peuvent être ensuite intégrés au cadre d'évaluation.

La question des effets non recherchés est également liée à celle de l'auteur de la théorie. Un bon protocole basé sur la théorie tiendra compte des théories concurrentes sur les modalités de fonctionnement d'un programme. Les responsables de programmes auront un point de vue, mais le personnel de terrain, les bénéficiaires et d'autres commentateurs peuvent en avoir d'autres. Par exemple, les responsables de programmes pensaient que les projets de fonds sociaux (fonds de développement décaissés au niveau des communautés) avaient des effets positifs sur le développement institutionnel local et national, à la fois par l'exemple (constater ce que faisait le fonds social) et par la pratique (lorsque ces agences participaient à la mise en œuvre du fonds social). Mais les critiques avançaient que les fonds sociaux contournaient les procédures publiques existantes, ce qui leur nuisait directement (en prenant du personnel) et moins directement en perturbant l'allocation optimale des ressources par les ministères de tutelle. L'évaluation considérait ainsi la théorie officielle du programme et l'anti-théorie concurrente (voir Banque mondiale, 2002, pour l'étude complète, Carvalho *et al.*, 2002 pour un résumé et Carvalho et White, 2004 pour une présentation de la démarche basée sur la théorie utilisée).

La documentation du projet est le point de départ habituel de l'élaboration de la théorie du programme. S'il y a un cadre logique, ce cadre représentera la théorie du programme. Cependant, il est rare que la documentation d'un projet explicite toutes les hypothèses sous-jacentes, même si certaines d'entre elles peuvent apparaître comme des « risques ». L'étape suivante consiste à présenter la théorie envisagée aux responsables du programme. Même s'ils n'y ont pas complètement réfléchi au préalable, ils auront un point de vue sur le document produit. Cet exercice est un bon moyen d'intéresser les responsables de programmes et de leur permettre d'influencer le protocole d'évaluation de manière positive[4]. La deuxième étape est de lire les études d'évaluation existantes et

[4] La réponse la plus fréquente des responsables de programmes est que le moment n'est pas bien choisi pour évaluer le programme parce qu'il vient juste d'être repensé, qu'ils viennent juste de faire leur propre étude, qu'il y a eu un changement de gouvernement, de ministre ou de chef de projet, etc. En général, il convient d'écarter poliment ces objections, comme bien sûr toute

la littérature théorique, s'il y en a, sur l'intervention évaluée ou sur des programmes comparables, qui identifieront les maillons faibles de la chaîne causale. Par exemple, l'erreur de ciblage est un problème souvent évoqué, surtout dans les programmes de microfinance (par exemple, Mosley et Hulme, 1996). Un point de vue plus nuancé est que la microfinance pour les femmes peut en fait être utilisée par les membres masculins du ménage, ce qui affecte l'impact sur les résultats finaux tels que la santé ou la nutrition des enfants. Il faut ensuite intégrer les points de vue des travailleurs de terrain et des bénéficiaires. Tout évaluateur pourra utilement se poser la question suivante : « Comment un villageois type vivra-t-il ce programme ? Comment en sera-t-il informé ? Pourquoi y participerait-il ? ». C'et exercice n'est pas inutile, même si l'anthropologie du développement nous a appris qu'en raison de perceptions différentes, de besoins différents ou d'un simple échec de communication de la part du personnel du projet, les points de vue locaux sur les projets peuvent être très différents de ce qui est anticipé.

Compréhension du contexte

La compréhension du contexte est indispensable pour appréhender l'impact du programme, et donc pour concevoir l'évaluation. Le contexte désigne le cadre social, politique et économique dans lequel s'inscrit le programme, tous ces éléments pouvant influencer le fonctionnement de la chaîne causale. Un même programme peut avoir un impact différent en fonction du contexte comme on l'a vu dans le cas du modèle TINP apparemment fructueux qui n'a pas très bien fonctionné au Bangladesh. En outre, comme il est dit plus loin, comprendre le contexte aide à anticiper l'hétérogénéité, et à généraliser.

Comprendre le contexte demande une lecture approfondie de la documentation du projet préalable à l'établissement du protocole d'évaluation, mais aussi une exposition à une littérature plus vaste (anthropologie et économie politique), comme nous le verrons plus loin dans la partie consacrée aux méthodes mixtes.

Une bonne compréhension du contexte facilite également la généralisation. Les études de l'appui apporté par la Banque mondiale à l'enseignement de base au Ghana et à la santé maternelle et infantile au Bangladesh concluaient globalement à la réussite des projets. Dans le cas du Ghana, des opérations à grande échelle de réhabilitation des écoles et de distribution de manuels scolaires avaient sensiblement amélioré le taux de scolarisation et les acquis scolaires (Banque mondiale, 2004). Deux éléments contextuels importants étaient à l'origine de ce résultat. D'une part, après des années de crise, le système scolaire était dans un piètre état ; les infrastructures étaient inadaptées et il n'y avait virtuellement aucune fourniture scolaire. Dans ce contexte, la rénovation des écoles et la distribution de manuels ont eu un impact qu'elles n'auraient peut-être pas eu si les écoles avaient déjà assez bien fonctionné. D'autre part, l'important soutien politique apporté à ce programme a facilité sa mise en œuvre (le programme faisait partie d'une

---

tentative d'influencer les constats. Mais il n'est pas inutile de relever ce que les responsables de programmes considèrent comme des questions d'évaluation importantes.

réforme de l'éducation). L'implication des pouvoirs publics a également été déterminante pour le succès du planning familial financé par l'aide qui a abouti à une transition démographique accélérée au Bangladesh, avec une chute spectaculaire de la mortalité et de la fécondité (Banque mondiale, 2005). En l'espace de dix ans, ce pays qui n'avait pratiquement aucun équipement après l'indépendance et que l'on donnait pour perdu lors de la famine qui a suivi, s'est doté d'un système national décentralisé de santé et de planning familial allant jusqu'à la fourniture à domicile de services de contraception. Des programmes aussi ambitieux peuvent vaciller si les pouvoirs publics ne sont pas déterminés à les mener à bien.

Anticipation de l'hétérogénéité

Une bonne compréhension du contexte aide à anticiper l'hétérogénéité potentielle de l'impact. L'impact (c'est-à-dire l'effet du traitement) peut en effet varier en fonction du protocole d'intervention, des caractéristiques des bénéficiaires ou du contexte socioéconomique. L'étude de la théorie sous-jacente peut aider à faire apparaître l'hétérogénéité possible et permettre au protocole d'évaluation de l'anticiper. Cette anticipation est importante pour deux raisons. Tout d'abord, les calculs de puissance pour la taille d'échantillon doivent refléter les niveaux de désagrégation qui seront utilisés dans l'analyse : plus le niveau de désagrégation est grand plus l'échantillon devra être important (pour les groupes de traitement et de contrôle). Deuxièmement, les lois de la probabilité suggèrent que si nous vérifions l'impact dans vingt sous-groupes définis arbitrairement, nous constaterons un impact significatif dans l'un d'entre eux au niveau de 5 %. Les bonnes pratiques imposées pour les ECR médicaux exigent de définir les sous-groupes à tester avant de recueillir les données. L'approche basée sur la théorie aide à pré-identifier ces groupes et donne une explication plausible au différentiel d'impact. Toutefois, il ne faut pas perdre de vue la nécessité d'itération entre modèle et données, un point sur lequel nous reviendrons plus loin.

Dans le cas de programmes d'alimentation des enfants, les enfants malnutris ont plus de chances de prendre du poids que les enfants déjà biens nourris, mais des enfants extrêmement malnutris peuvent souffrir de diarrhées qui empêchent une alimentation efficace et une prise de poids. Des programmes mieux ciblés auront donc un impact moyen plus élevé et l'impact sera plus net à la période de soudure – ce qui a été effectivement constaté dans le cas du BINP. Ce sont sans doute les enfants les plus jeunes qui bénéficieront le plus du programme ; pour ceux qui ont souffert d'un retard de croissance dans leur petite enfance, l'alimentation donnée ultérieurement ne permettra pas de gains de taille considérables. De même, les gains cognitifs résultant d'une meilleure nutrition semblent acquis à moins de trois ans. L'impact varie donc en fonction de l'âge des bénéficiaires et de l'état nutritionnel préexistant, ce dernier ayant une composante de saisonnalité. L'impact peut également varier en fonction du statut socioéconomique ; la substitution (le remplacement d'un repas par le supplément alimentaire) est plus probable dans les ménages pauvres par exemple.

C'est pour ces raisons que les programmes alimentaires, qui reposaient auparavant sur l'alimentation à l'école, tendent désormais à cibler les enfants de moins de trois ans, comme dans l'exemple du Bangladesh évoqué plus haut. Cependant, l'alimentation à l'école peut encore produire des gains d'apprentissage. Les enfants qui souffrent de carences caloriques sont fatigués et manquent d'énergie ; un programme d'alimentation peut donc les rendre plus attentifs en classe, à ceci près qu'un bon repas induit souvent une somnolence, si bien que le moment est important. Mais le cadre est important pour que les enfants les plus attentifs enregistrent eux aussi des gains d'apprentissage. Pour toute intervention, il est crucial de s'attaquer à la bonne contrainte. Il ne sert à rien à un enfant d'être attentif si l'enseignant est absent, et l'enfant apprendra sans doute moins s'il n'a pas de matériel pédagogique. On peut donc penser que l'impact des programmes alimentaires sera plus marqué dans les établissements scolaires qui fonctionnent bien que dans les écoles mal équipées où l'absentéisme des enseignants est fréquent. Une observation similaire a été faite concernant les allocations sociales conditionnelles, qui accroissent la demande de scolarisation mais n'améliorent pas nécessairement les acquis de l'enseignement ni même le taux de scolarisation s'il existe des contraintes du côté de l'offre (Ravaillon, 2009).

La complémentarité possible des interventions est un autre aspect de l'hétérogénéité. Ainsi, la microfinance a un impact large si elle est accompagnée de services d'appui aux entreprises, mais il est possible que les deux interventions soient des substituts, leur impact combiné étant inférieur à la somme des deux interventions séparées. Il est clair que les protocoles d'évaluation qui examinent ces complémentarités ont une grande pertinence politique.

L'impact peut varier dans le temps, malgré l'hypothèse habituelle (souvent implicite) de linéarité de sa trajectoire (Woolcock, 2009). Linéarité de la trajectoire d'impact ne veut pas dire causalité unidirectionnelle, dont on a vu plus haut qu'elle est critiquée, ni approche statique face à la dynamique de la mise en œuvre du programme. Même lorsque le schéma du programme reste inchangé et que la direction causale a été établie, les impacts de l'intervention peuvent varier dans le temps et les constats seront très sensibles au moment auquel l'impact est mesuré. Dans le cas de projets destinés à développer la participation et l'autonomie de groupes marginalisés par exemple, la littérature sociale suggère qu'« en fait, la forme fonctionnelle la plus probable de ces projets est une courbe en J, c'est-à-dire que la situation empire avant de s'améliorer – du moins l'espère-t-on ». C'est un aspect qui n'a pas été suffisamment étudié en utilisant l'évaluation d'impact basée sur la théorie, mais qui s'y prête particulièrement bien. Le programme BINP évoqué plus haut a pu induire des conflits initiaux entre les femmes et leurs maris et belles-mères du fait de la sensibilisation accrue des femmes, ce qui n'expliquerait aucun impact nutritionnel, mais il est possible qu'une évaluation à plus long terme décèle un effet positif compte tenu de l'évolution sociale plus vaste qui accroît le statut des femmes dans le Bangladesh rural.

Le repérage de l'hétérogénéité est lié à la capacité de généralisation. Un ECR au Kenya, en Afrique du Sud et en Ouganda a vérifié l'impact de la circoncision des hommes sur la

transmission du VIH/SIDA et constaté que les hommes circoncis avaient une bien plus faible propension à contracter la maladie (voir par exemple Wawer *et al.*, 2008, sur l'Ouganda). L'âge était un des aspects de l'hétérogénéité. La circoncision doit être suivie d'un mois d'abstinence pour que la plaie puisse cicatriser ; les rapports sexuels dans cette période sont plus risqués, et non pas moins risqués. Cette opération réalisée sur des garçons de 12 ans par exemple n'engendre pas le risque d'une exposition d'un mois à un risque élevé, tandis que les hommes adultes sont souvent incapables de s'abstenir pendant un mois entier, ce qui réduit l'impact bénéfique du traitement. Pourtant, les études ont constaté que la circoncision diminuait le risque de transmission de 30 à 50 %. Ce niveau d'impact ne peut être généralisé qu'aux populations dont le comportement sexuel est comparable. Dans une communauté au sein de laquelle les hommes pratiqueraient l'abstinence, n'auraient qu'une partenaire sexuelle ou utiliseraient tous le préservatif, l'intervention n'aurait pas d'impact.

Impact

Bien entendu, une évaluation rigoureuse de l'impact faisant appel à un contrefactuel approprié est un élément clé de l'évaluation d'impact basée sur la théorie. Le contrefactuel approprié est le plus souvent défini par rapport à un groupe de contrôle, qui doit être déterminé de manière à éviter tout biais de sélection, c'est-à-dire en faisant appel à des méthodes expérimentales ou quasi expérimentales. Disposer de données de panel renforce le protocole ; il faut donc encourager les enquêtes de référence – conçues de manière à permettre la ré-identification des ménages de l'échantillon. Lorsqu'on n'en dispose pas, il sera peut-être possible d'en recréer à partir d'ensembles de données existants ou de souvenirs, même si cette dernière solution invite à la prudence (voir Bamberger, 2009). Outre le biais de sélection, les autres aspects importants à envisager dans le protocole sont la possibilité d'effets de diffusion (le groupe de contrôle est affecté par l'intervention) et de contagion (le groupe de contrôle est affecté par d'autres interventions).

Analyse factuelle rigoureuse

L'analyse contrefactuelle de l'impact doit être complétée par une analyse factuelle rigoureuse. Nombre des liens de la chaîne causale reposent sur l'analyse des faits. Dans le cas du BINP, celle-ci comprenait un mauvais ciblage et les raisons de celui-ci, l'identification des déperditions et le fait que les connaissances acquises n'aient pas été mises en pratique.

L'analyse de ciblage est la forme la plus courante d'analyse factuelle ; elle doit faire partie intégrante de la majorité des études d'impact, voire de toutes : qui bénéficie du programme ? Dans la mesure où un groupe cible est défini, quelle est l'ampleur des erreurs de ciblage ; ces erreurs peuvent être quantifiées et leur source identifiée, comme cela a été fait au Bangladesh. L'analyse de ciblage doit être effectuée à différents niveaux. Dans le cas des fonds sociaux, on a constaté que dans de nombreux pays, l'utilisation de cartes de la pauvreté conduisait à cibler les fonds sociaux sur les régions

les plus pauvres, mais que dans ces régions, c'étaient les communautés les mieux loties qui avaient le plus de chances d'accéder aux ressources des projets (Banque mondiale, 2002). De même, dans le cas de l'électrification rurale, les communautés les mieux loties avaient plus de chances de se raccorder, mais les ménages les plus pauvres des communautés raccordées au réseau restent sans électricité pendant plusieurs années parce qu'ils n'ont pas les moyens de payer les frais de raccordement (Banque mondiale, 2008).

L'analyse de ciblage doit être effectuée avec un ensemble de données représentatif. Lorsque les échantillons laissent la possibilité de biais de sélection, il arrive souvent que les données de l'évaluation d'impact ne soient pas représentatives de l'ensemble de la population et qu'elles ne puissent pas répondre à une question telle que « quel pourcentage des 20 % les plus pauvres bénéficient du projet ? » sauf si des poids d'échantillonnage permettent de rendre l'échantillon représentatif.

La deuxième remarque concernant le ciblage est que c'est un exercice portant sur deux variables, qui consiste à représenter la participation sous forme de graphique ou de tableau par rapport aux caractéristiques étudiées (celles-ci pouvant être celles des individus, des ménages ou d'une communauté). Une approche quasi expérimentale requiert une analyse multivariée de la participation au programme, mais c'est généralement une erreur d'utiliser ces résultats pour l'analyse de ciblage, car celle-ci doit reposer sur des statistiques descriptives. Le fait que le programme atteigne ou non les 20 % inférieurs est une observation reposant sur un tableau à deux variables, et non l'importance statistique d'un quintile dans une régression multivariée. La régression peut mettre les facteurs de la participation en évidence et aider ainsi à expliquer les résultats bivariés du ciblage. Par exemple, une analyse multivariée d'un projet en Inde pourrait révéler une participation beaucoup plus faible des populations tribales, qui sont parmi les plus pauvres de certaines régions du programme, ce qui explique les insuffisances du ciblage[5].

L'opération consistant à vérifier si les personnes qui ont été exposées à une formation ont appris, et mis en pratique, les nouvelles approches recherchées est un exemple de forme sous-utilisée d'analyse factuelle. L'étude BINP a montré que les mères acquéraient des connaissances mais que beaucoup ne les mettaient pas en pratique. Et les agents communautaires de nutrition pouvaient effectuer des séances de pesée, mais et c'est là le plus important, elles n'avaient pas appris à interpréter correctement les courbes de croissance. Ce type d'analyse n'est pas souvent pratiqué, mais il est clair qu'elle offre d'importantes possibilités. Les enseignants formés connaissent-ils les méthodes

[5] Plus précisément, un terme richesse est significatif lorsqu'une variable muette d'appartenance tribale est exclue, mais devient négligeable lorsqu'on inclut cette variable. La participation est donc déterminée par l'appartenance tribale, plutôt plus que par la pauvreté en soi. Cette démarche n'est pas toujours possible du fait du degré élevé de colinéarité entre les variables explicatives possibles, telles que celles mentionnées ainsi que l'éducation et la localisation.

pédagogiques améliorées et les mettent-ils en pratique ? Une étude de la Banque mondiale au Ghana indiquait que beaucoup ne le font pas[6].

Comme dans le cas du Bangladesh, l'analyse factuelle peut souvent mettre en évidence une rupture cruciale dans la chaîne causale et expliquer ainsi la faiblesse d'un impact. Une autre étude de l'OED a constaté que les services de vulgarisation agricole au Kenya n'avaient aucun impact sur les rendements. En principe, le projet finançait de nouvelles recherches agricoles dans des stations de recherche, dont les enseignements étaient transmis à des vulgarisateurs, puis aux agriculteurs. En réalité, les enseignements de la recherche n'étaient pas transmis aux vulgarisateurs, qui invitaient les agriculteurs à adopter des pratiques que la plupart avaient déjà adoptées depuis longtemps (Banque mondiale, 2000).

Cependant, il arrive que le besoin apparent d'analyse factuelle soit en fait un besoin de contrefactuel. Une subvention scolaire proportionnelle au nombre d'élèves est destinée à augmenter le taux de scolarisation et les résultats de l'enseignement, mais comment le fait-elle ? L'explication doit résider dans l'emploi qui est fait des fonds. Cette analyse peut ressembler à une analyse factuelle classique consistant à retracer l'emploi des fonds : vérifier quel montant parvient effectivement aux écoles et comment il est dépensé, des éléments utiles à l'étude. Toutefois, si les écoles ont déjà des ressources à disposition, il y a la possibilité de la fongibilité. Dans ce cas, une analyse avant-après des dépenses pourrait produire un contrefactuel valide, même si une analyse des améliorations de l'école et de l'acquisition de matériels dans le groupe de traitement et dans le groupe de contrôle sera sans doute plus productive.

Mixité des méthodes

La mixité des méthodes consiste à associer des méthodes quantitatives et qualitatives dans une seule évaluation, sachant que toutes les études quantitatives ont un élément qualitatif – au moins la lecture de la documentation de projet ; c'est donc une affaire de degré.

En règle générale, ce sont les partisans des approches qualitatives qui plaident pour les méthodes mixtes. Mais jusque récemment, ce sont les approches qualitatives qui ont dominé l'évaluation en matière de développement ; l'adoption de méthodes mixtes consiste donc à utiliser davantage de méthodes quantitatives rigoureuses dans les études qualitatives. Mais je parle ici de développer l'utilisation des données qualitatives dans les études quantitatives, un point que j'ai traité plus amplement dans White (2008). Je ferai trois remarques générales.

[6] L'observation de classe serait le meilleur moyen de mesurer les pratiques, mais cette méthode a été exclue pour des raisons de coûts. On peut penser qu'interroger les enseignants sur leurs méthodes produirait des résultats biaisés car ils déclareraient appliquer de meilleures méthodes, même s'ils ne le font pas.

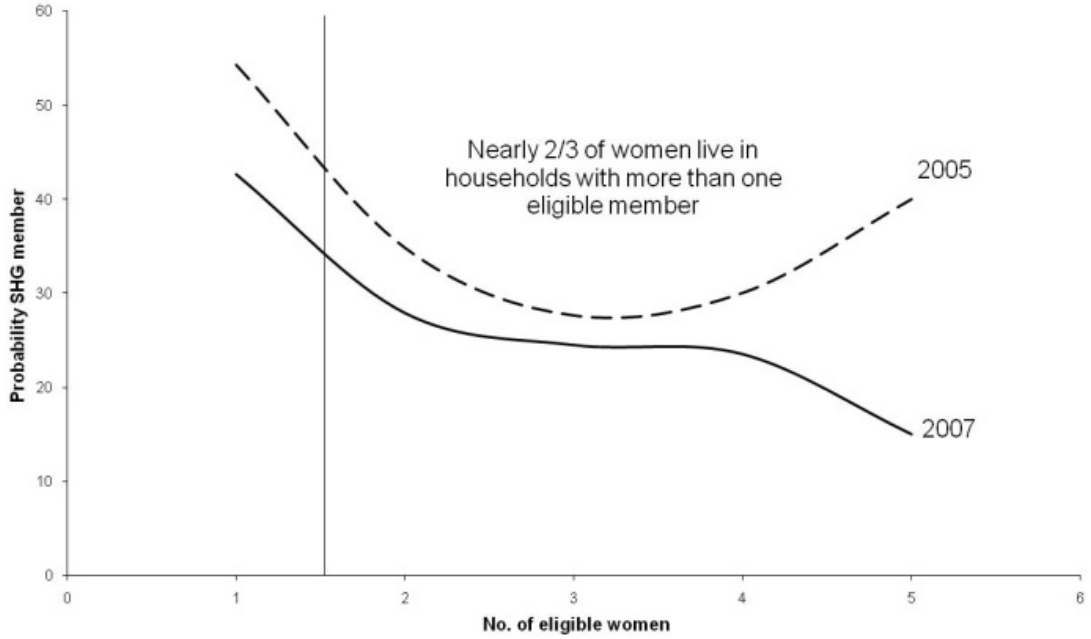
Premièrement, l'utilisation de données qualitatives implique un large éventail d'activités, qui ne se limite pas à l'organisation de focus groups (lesquels font à mon avis partie des formes de données qualitatives les moins fiables, sauf si elles sont vraiment bien conduites). Cela suppose par exemple de lire la littérature anthropologique et politique du contexte d'intervention pour éclairer le protocole d'évaluation. Dans le cas du Bangladesh, le repérage de l'effet « belle-mère » est venu de la lecture d'ouvrages d'anthropologie (en particulier White, 1992). Cet éclairage nous a conduit à analyser la partie du questionnaire sur la liste des ménages pour identifier les femmes qui vivaient avec leur belle-mère (par exemple, belle-fille du chef de famille dont l'épouse est également présente, épouse du chef de famille dont la mère est également présente et belle-sœur du chef de famille dont la mère est également présente) et exécuter ainsi une analyse quantitative éclairée par des données qualitatives.

L'éventail des techniques va du « tourisme de développement » (passer une journée ou deux sur le terrain), à la boîte à outils de l'évaluation rurale participative (ERP), jusqu'à l'intégration d'un anthropologue dans la zone de projet, cette dernière solution étant une démarche sous-exploitée qui pourrait être utilisée pour des études à plus long terme. Ma deuxième remarque est que si décrié qu'il soit, le tourisme de développement est un pan essentiel de l'évaluation d'impact basée sur la théorie. Il n'y a réellement aucun substitut au temps que vous passez vous-même sur le terrain, et on peine à voir comment les données peuvent être intelligemment analysées en l'absence d'exposition au terrain (ça se voit quand c'est le cas). Quelques jours d'exposition à la mise en œuvre du projet dans un ensemble de contextes – et de préférence sans se limiter à ceux qui ont été choisis par le personnel du projet – aideront à concevoir l'étude et à la mettre en œuvre.

Je pourrais donner de nombreux exemples des éclairages apportés par des discussions avec le personnel du projet, les bénéficiaires et d'autres parties prenantes sur le terrain. Je n'en donnerai que deux. Le premier est tiré d'une évaluation d'un projet de moyens d'existence en milieu rural, dans le cadre duquel des prêts étaient consentis par l'intermédiaire de groupes d'entraide féminins. Un homme se plaignait que sa fille non mariée de 22 ans ne pouvait pas obtenir de prêt parce que sa femme en avait déjà obtenu un. Cette remarque a permis de prendre conscience que les villageois pensaient que le prêt était octroyé au ménage et non à l'individu, un fait qui explique les taux de participation bien plus faibles des femmes dans les ménages qui comptent plus d'une femme remplissant les conditions requises pour participer à un groupe d'entraide (Figure 2). L'objectif du projet était que toutes les femmes admissibles participent, mais ce n'est pas un objectif réaliste tant que les bénéfices sont au niveau du foyer et non des individus. Le deuxième exemple montre la force de généralisation que peut avoir une formule bien choisie. Dans le travail de terrain pour l'évaluation du fonds social zambien, il était frappant que tous – des responsables au personnel du programme en passant par les villageois – disaient « la communauté » a choisi le projet, même s'il était évident qu'un processus bien plus sélectif était à l'œuvre (voir White et Vajja, 2008, pour une analyse plus longue). Cependant, le fait que « la communauté » était en fait un concept assez étroit, qui désignait en réalité le comité de projet, a été bien saisi par un responsable de programme régional qui, répondant à un appel sur son téléphone

portable, nous déclara « Il faut que je raccroche, j'ai une communauté dans mon bureau ».

Figure 2 – Taux de participation aux programmes au sein de groupes d'entraide dans l'Andhra Pradesh par nombre de femmes admissibles dans le ménage



Source : données d'enquête IEG

Probabilité de participation à un groupe d'entraide

Près de 2/3 des femmes vivent dans des ménages comptant plus d'un membre admissible

Nombre de femmes admissibles

Puisqu'on dispose de données permettant d'évaluer l'impact sans recueillir de nouvelles données (ce qu'il faut encourager car on a trop tendance à collecter des données alors que les données dont on dispose sont sous-exploitées), le danger existe que les chercheurs conduisent des évaluations d'impact sans aucune exposition à l'intervention. Ces études ont toutes chances de manquer de pertinence pour les politiques du fait d'une compréhension insuffisante des mécanismes réels de fonctionnement de l'intervention.

Enfin, le budget doit permettre des activités de type recherche-action, où les problèmes présentés par les données sont suivis d'un travail de terrain complémentaire. Les focus groups sur les raisons de l'écart entre théorie et pratique au Bangladesh sont un exemple de ce type de travail. Un autre exemple vient d'une étude de financement de groupes d'entraide en Andhra Pradesh en Inde. Nous avions des données de panel et l'étude comprenait un module standard de type enquête sur la mesure des niveaux de vie sur les entreprises familiales. L'analyse de ces données a montré que la plupart des entreprises étaient peu rentables et qu'une minorité étaient déficitaires. Mais les données du modèle

étaient un instrument vraiment trop grossier pour comprendre comment ces entreprises fonctionnaient. Nous avons donc recouru à ce que j'appellerai « l'éthnographie quantitative » pour revoir tous les ménages qui avaient été interrogés et déclarés comme étant une entreprise. Cette deuxième visite utilisait un questionnaire semi-structuré pour identifier le cash-flow journalier des entreprises et les apports de travail des membres du ménage (et des employés, même si ceux-ci étaient très rares). Les résultats ont effectivement confirmé le faible niveau de revenu de ces activités (20-30 INR par jour n'était pas rare, à comparer avec un salaire journalier de 50-70 INR), et la nature risquée de certaines (mort des animaux, surtout des chèvres, et taille de marché insuffisante).

4. Évaluation d'impact basée sur la théorie et approches de type boîte noire

On peut contraster l'évaluation d'impact basée sur la théorie avec une approche de type « boîte noire ». Cette dernière ne fait bien souvent que rapporter un impact – car elle s'intéresse à la signification statistique du coefficient pour l'effet moyen du traitement, mais elle ne tente en aucun cas de répondre à la question du pourquoi. Cet article a voulu montrer comment aborder la question du pourquoi et l'intérêt d'une telle démarche. Cependant, quelques avertissements s'imposent.

Il ne faut pas exagérer les critiques relatives à l'effet moyen du traitement. L'hétérogénéité est importante, tout comme la compréhension du contexte dans lequel un impact particulier s'est produit, mais il est rare que l'effet moyen du traitement (habituellement en traitement des traités et en intention de traiter) n'ait aucun intérêt. En fait, il est très probable que ce soit le principal paramètre d'intérêt. Il serait trompeur de faire état d'une signification ou de l'absence de signification pour un sous-groupe particulier si l'effet moyen du traitement avait le signe opposé. De plus, l'effet moyen du traitement sert de base aux calculs de coût-efficacité.

Deuxièmement, l'évaluation d'impact basée sur la théorie explicite la chaîne causale de diverses manières. Elle s'efforce de démêler ses différentes étapes, mais aussi de faire la part entre les éléments d'une intervention qui fonctionnent et ceux qui ne fonctionnent pas. On peut pour cela procéder à une analyse de régression. L'étude du BINP, par exemple, présente des régressions sur les déterminants de l'écart entre connaissances et pratiques. Mais ces approches basées sur la régression, qui s'appuient sur des modèles de sélection d'échantillons et la spécification paramétrique de la relation examinée, ont de nombreux critiques, qui préfèrent les approches expérimentales ou quasi expérimentales telles que le PSM et le RDD. Ces approches rigoureuses peuvent accueillir une analyse des éléments du programme qui fonctionnent, mais l'intervention doit être conçue de façon à pouvoir moduler le protocole d'intervention d'un groupe à l'autre – par exemple, certains entrepreneurs obtiennent des prêts, d'autres obtiennent des services d'aide aux entreprises, et d'autres obtiennent les deux. En pratique, une évaluation

d'impact basée sur la théorie combinera les estimations d'impact rigoureuses qui peuvent être effectuées avec d'autres approches d'explicitation de la chaîne causale.

Enfin, ce qui est à l'intérieur de la boîte noire peut être si confus qu'il vaut mieux parfois ne pas soulever le couvercle. L'étude de l'électrification rurale conduite par la Banque mondiale examinait l'impact de l'électrification sur la fécondité. L'accès à l'électricité réduit sensiblement la fécondité (Banque mondiale, 2008). L'étude a pu démontrer un canal possible qui semblait en jeu (l'accès à la télévision augmentant la connaissance sur la contraception) et un canal qui ne l'était pas (les occupations non sexuelles réduisant l'activité sexuelle). Mais il peut y avoir d'autres canaux, tels que les effets de revenu, d'autres avantages éducatifs, etc., où il est impossible de séparer tous les canaux ; dans ce cas, une forme réduite d'estimation d'impact peut être la meilleure solution.

5. Conclusions

Cet article souscrit aux appels à produire un plus grand volume d'études quantitatives rigoureuses sur les interventions fructueuses en matière de développement. Cependant, ces études seront bien plus pertinentes pour les politiques publiques si elles apportent un éclairage sur les raisons pour lesquelles les interventions marchent ou ne marchent pas. On s'accorde généralement à penser que l'évaluation d'impact basée sur la théorie peut produire les éclairages nécessaires. Toutefois, de nombreuses études nouvelles ne tiennent pas la promesse de l'approche basée sur la théorie, car elles émettent des hypothèses sur les raisons de l'impact ou les différences d'impact au lieu de les expliquer par une analyse empirique solide.

J'ai présenté un exemple pratique d'évaluation d'impact basée sur la théorie et montré que cette méthode conduit directement à des conclusions politiques pour renforcer l'impact du programme. Pour cela, il a fallu appliquer les principes exposés plus haut. La théorie du programme doit être flexible pour pouvoir s'adapter à l'évolution des circonstances sur le terrain et accueillir des théories concurrentes et des conséquences involontaires. La rigueur doit être combinée dans une analyse factuelle et contrefactuelle, ce qui suppose d'employer une combinaison de méthodes. La théorie du programme doit être établie dans le contexte social, politique et culturel de l'intervention, ce qui permettra de mettre en lumière l'hétérogénéité attendue de l'impact.

Références

3ie, Initiative internationale pour l'évaluation d'impact (non daté), « 3ie impact evaluation practice: a guide for grantees », http://www.3ieimpact.org/page.php?pg=overview (accès le 1er juin 2009).

Bamberger, Michael (2009), « Strengthening the evaluation of program effectiveness through reconstructing baseline data », *Journal of Development Effectiveness* **1**(1): 37-59.

Banque mondiale (2000), *Agricultural extension: the Kenya experice*, OED, Banque mondiale, Washington D.C.

Banque mondiale (2002), *Social Funds, assessing effectiveness,* OED, Banque mondiale, Washington D.C.

Banque mondiale (2005), *Maintaining Momentum to 2015? An impact evaluation of interventions to improve maternal and child health and nutrition in Bangladesh,* OED, Banque mondiale, Washington D.C.

Banque mondiale (2006), *Repositioning Nutrition as Central to Development: a strategy for long term large-scale action,* Banque mondiale, Washington D.C.

Banque mondiale (2008), *The welfare impact of rural electrification: a re-assessment of the costs and benefits,* IEG, Banque mondiale, Washington D.C.

Blackman, Leonard et Stephanie Reich (2009), « Randomized control trials: a gold standard with feet of clay? », Stewart Donaldson, Christina Christie et Melvin Mark (dir. pub.) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Sage, Thousand Oaks, Californie.

Carvalho, Soniya, Gil Perkins et Howard White (2004), « Social funds: participation, social capital and sustainability », *Journal of International Development,* vol. 14, p. 611-625, 2002.

Carvalho, Soniya et Howard White (2004), « Theory-based evaluation: the case of social funds », *American Journal of Évaluation*, vol. 25, n° 2, p. 141-60, 2004.

Centre for Global Development (2006), *When Will We Ever Learn?,* Centre for Global Development, Washington D.C..

Leamer, E. (1983), « Let's take the con out of econometrics », *American Economic Review*, vol. 23, n° 1, 31–43.

Mosley, Paul et David Hulme (1996) *Finance Against Poverty,* Routledge, Londres.

Mukherjee, Chandan, Marc Wuyts et Howard White (1994) *Econometrics and Data Analysis for Developing Countries*, Routledge, Londres.

NONIE (non daté) « NONIE statement on impact evaluation » http://www.worldbank.org/ieg/nonie/members.html (accès le 1er juin 2009).

Rogers, Patricia (2009), « Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation », *Journal of Development Effectiveness* **1**(3).

Save the Children (2003), *Thin on the Ground. Questioning the evidence behind Banque mondiale-funded community nutrition projects in Bangladesh, Ethiopia and Uganda*. Save the Children UK, Londres.

Wawer M, Kigozi G, Serwadda D, *et al.,* « Trial of Male Circumcision in HIV+ Men, Rakai, Uganda: Effects in HIV+ Men and in Women Partners », 15e conférence sur les rétrovirus et les infections opportunistes, 2008, Boston, MA, 2008.

Weiss, Carol (1998), *Evaluation: methods for studying programs and policies*. Prentice Hall, New York.

White, Howard (2005), « Comment on Contributions Regarding the Impact of the Bangladesh Integrated Nutrition Project », *Health Policy and Planning*, vol. 20, n° 6, p. 408-411.

White, Howard (2008), « Of Probits and Participation: the use of mixed methods in quantitative impact evaluation » *IDS Bulletin*, 2008.

White, Howard (2009), « Some reflections on current debates in impact evaluation », *3ie Working Paper No. 1*, Initiative internationale pour l'évaluation d'impact, New Delhi.

White, Howard et Edoardo Masset (2006), « The Bangladesh Integrated Nutrition Program: findings from an impact evaluation », *Journal of International Development*, vol. 19, p. 627-652, 2006.

White, Howard et Anju Vajja (2008), « Can the World Bank Build Social Capital?: Community Participation in Social Funds in Malawi and Zambia », *Journal of Development Studies*, vol. 33, n°8, p. 1145-1168.

White, Sarah (1992) *Arguing with the crocodile: gender and class in Bangladesh*, Zed, Londres.

Woolcock, Michael (2009) « Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy », *Journal of Development Effectiveness*, vol. 1, n° 1, p. 1-14.