

La randomisation à l'épreuve du partenariat : l'exemple du projet de microassurance SKY au Cambodge.

Aurélie Quentin (Gret), Isabelle Guérin (UMR 201 Développement et Société (Paris I Sorbonne / IRD) et Laura Perez Medina (Gret).

Draft

I. INTRODUCTION

Depuis une dizaine d'années, on assiste à un fort engouement des bailleurs internationaux pour les études quantitatives de type « randomisées »¹ visant à évaluer l'efficacité des projets ou politiques de développement qu'ils financent. La mise en avant des études expérimentales par assignation aléatoire s'inscrit dans la montée en puissance au niveau international d'un paradigme de prise de décision politique fondée sur des preuves scientifiques (« evidence based policy »). Cette approche, critiquée par certains pour ses effets en terme de dépolitisation et saluée par d'autres comme un progrès empirique, est activement promue par la Banque Mondiale (Bedecarrats, 2011). Cela tend à éclipser voire à discréditer les autres types d'approches visant à évaluer l'impact des programmes de développement, qu'elles soient quantitative, qualitative ou les deux à la fois.

Or il est aujourd'hui reconnu de manière assez large qu'il est souhaitable d'arriver à concilier, dans le cadre de ces évaluations, d'une part les préoccupations de rigueur statistique visant à mesurer et démontrer (« prove ») les impacts effectifs des politiques et programmes mis en œuvre ; d'autre part les préoccupations de compréhension des phénomènes et logiques des acteurs concernés visant plutôt à améliorer et adapter les dispositifs et méthodes d'intervention (« improve ») des opérateurs de développement [Bouquet et al, 2009]. Il y a un enjeu scientifique autour de l'utilisation conjointe et synergique de méthodes quantitatives et qualitatives pour améliorer la compréhension et l'analyse des phénomènes observés, outils eux-mêmes mobilisés préférentiellement par des disciplines différentes, économie d'un côté, socio-anthropologie de l'autre.

Depuis 1998, le Gret - Ong française de développement adoptant une approche fondée sur le partenariat et la recherche-action - développe un projet de micro-assurance santé au Cambodge, le projet SKY. Après une phase pilote initiale, celui-ci a été étendu en 2004, avec l'appui financier de l'Agence Française de Développement (AFD).

Au milieu des années 2000, le Département Recherche de l'AFD a souhaité mettre en place un dispositif innovant d'évaluation d'impact sur ce programme, visant à mesurer par des méthodes à la fois quantitative, notamment à travers l'approche expérimentale, et qualitative, l'impact du programme sur la santé des populations cibles. Dans cet objectif, l'AFD s'est

¹ « L'expérimentation par assignation aléatoire est une technique issue des sciences de la santé et importée depuis une douzaine d'années environ dans le champ des politiques publiques et de l'économie du développement. » [Labrousse, 2010]

adressé au CEGA (Center of Evaluation for Global Action - UC Berkeley)² pour formaliser un accord de financement en 2007.

En 2011, la direction scientifique du Gret et l'IRD ont initié une recherche ayant pour objet de documenter et d'analyser le processus de mise en place et le déroulement même de l'étude d'impact du programme SKY, et en particulier la manière dont ont pu être pris en compte les préoccupations et intérêts parfois divergents des parties prenantes, et d'en analyser les premiers résultats. Cette étude menée entre mai et décembre 2011 - s'appuyant sur le recrutement d'une stagiaire de master 2 et comprenant un séjour de trois mois au Cambodge entre juillet et octobre - a consisté à retracer le déroulement de l'évaluation en analysant l'ensemble des documents formels et informels produits dans ce cadre, en réalisant des entretiens avec l'ensemble de ses parties prenantes, et en étudiant une grande partie des e-mails échangés entre 2006 et 2011 par ses principaux protagonistes, matériau fourni à la fois par la responsable chargée du suivi de l'étude d'impact au sein de la cellule évaluation de l'AFD et par le représentant du GRET au Cambodge responsable du projet SKY sur la période.

L'objectif de notre recherche est d'interroger les modalités concrètes d'articulation et de compromis entre les contraintes de la rigueur académique et les exigences opérationnelles, ainsi qu'entre différents modes de production des savoirs. Plus généralement, il s'agit d'apporter, sur la base d'éléments empiriques, des éléments de réponse aux questions d'ordre politique, éthique et méthodologique posées par cette méthode d'évaluation.

Pour cela, l'analyse menée comporte un double niveau :

- elle porte d'une part sur les contraintes classiques posées par le dialogue entre méthodes et entre chercheurs/opérateurs ;
- elle s'appuie d'autre part sur le décryptage d'un processus particulier d'évaluation à travers un examen des jeux d'acteurs, des compromis, des négociations mais aussi des tensions et des conflits qu'il suscite, des rapports (très inégaux) de pouvoir, des systèmes de contraintes et des enjeux de l'évaluation pour chaque type d'acteur (opérateurs, évaluateurs, et bailleurs de l'évaluation).

Par rapport à l'objectif officiel de départ – évaluer l'impact du projet SKY et en tirer des leçons en matière de politique publique de santé – cette étude est de toute évidence un échec. Les différentes parties prenantes de l'étude n'ont finalement jamais réussi à s'entendre sur un objectif commun. Les contraintes méthodologiques ont non seulement considérablement restreint les questions posées mais ont également conduit à déformer le projet étudié afin qu'il puisse se prêter aux exigences techniques de la méthode. La validité externe de l'étude, c'est-à-dire les conclusions opérationnelles et politiques que l'on peut en tirer, sont donc très limitées.

Au-delà des résultats proprement dit, cette étude questionne la capacité réelle des études randomisées à faire l'objet de processus négociés entre acteurs : le poids des contraintes de méthode limite considérablement la marge de négociabilité concernant les objectifs et le spectre des questions possibles. En dépit de tentatives répétées, l'opérateur a très peu réussi à influencer les orientations de l'étude. Sa principale marge de manœuvre a consisté

² Au sein de cet article, l'acronyme CEGA pourra également être orthographié CEGA. En effet, à l'origine il s'agissait d'un programme de recherche intitulé *Scientific Evaluation of Global Action* qui est devenu à la fin des années 2000 un centre de recherches de Berkeley sous le nom de *Center of Evaluation for Global Action*.

à en limiter les perturbations sur ses activités opérationnelles, en exigeant notamment une certaine forme d'échantillonnage. Or exigences posées ont dans une certaine mesure perturbé le principe même de la méthode aléatoire, ce qui questionne la validité interne de l'étude.

II. LA REALISATION DE L'ETUDE D'IMPACT ET SON RAPPORT A L'OPERATIONNEL : UN DISPOSITIF PLURI-ACTEURS EN PERPETUELLE NEGOCIATION.

1. Les origines du projet.

1.1 Une injonction sur la méthode qui précède le choix de l'objet d'étude

En décembre 2005, en marge de la conférence annuelle AFD-EUDN qui avait alors pour thème le financement de l'aide au développement, Ester Duflo (J-PAL - MIT) et David Levine (Haas – UC Berkeley), membres du « *evaluation gap working group* » du Center for Global Development (CGDev) animent des conférences sur l'évaluation d'impact. Leur objectif est de promouvoir à très haut niveau, auprès notamment des directions générales des différentes agences d'aide internationale, le financement et la pratique d'évaluation « rigoureuses » d'impact des programmes de développement mis en œuvre. Le rapport emblématique de ce groupe de travail, intitulé « *when will we ever learn* » et publié en mai 2006, a un à très fort retentissement dans le monde du développement, dans le contexte d'un renouvellement du débat sur l'efficacité de l'aide à la suite de la déclaration de Paris en 2005.

C'est dans ce cadre que D. Levine et le département évaluation de l'AFD conviennent de travailler ensemble sur la mise en œuvre d'une évaluation d'impact de type randomisée (RCT). Cette décision se fonde sur plusieurs éléments : d'une part, dans un contexte de fort engouement pour cette « nouvelle » méthode³, défendue et diffusée à l'échelle mondiale notamment par les travaux d'Ester Duflo, et de fort lobbying des différents acteurs liés au CGDev – dont le groupe de travail sur l'évaluation avait été financé par la fondation Bill et Melinda Gates avec une participation active des économistes de la Banque Mondiale –, la direction générale de l'AFD souhaite financer et suivre de très près des évaluations qui l'appliquent à certains de ses programmes. Le choix est de nature stratégique : plutôt que de simplement contribuer financièrement à de grands fonds mondiaux tels que le 3IE mis en place par la suite par le CGDev pour canaliser les contributions de différentes institutions au financement des RCT, l'AFD souhaite internaliser cette compétence en se forgeant une expertise en la matière tout en évaluant concrètement les apports et les limites de cette approche. D'autre part, le positionnement spécifique de l'AFD comme bailleur ayant un rôle actif en matière de production de connaissances établissant des partenariats scientifiques internationaux, conduit l'agence à re-

³ La méthode n'a en elle-même rien de nouveau puisque les études expérimentales en sciences sociales ont connu un fort développement entre la fin des années 1960 et le début des années 1980 notamment aux Etats-Unis, ce qui lui confère un caractère novateur est son application à l'évaluation de projets de développement, et donc à des contextes sociaux et culturels moins maîtrisés par les chercheurs qui l'appliquent.

chercher des collaborations avec des universités prestigieuses telles que Berkeley ou le MIT, et à se positionner sur des sujets importants pour le monde académique.

Le département évaluation de la division recherche de l'AFD est donc fortement incité par la direction générale à établir des partenariats scientifiques, financer et tirer des enseignements de la supervision d'études d'impacts randomisées, ce qui conduit au lancement de deux grandes études en 2006, l'une sur la microfinance en milieu rural au Maroc réalisée par une équipe du J-PAL et de l'Ecole d'Economie de Paris (EEP), l'autre sur la micro-assurance santé au Cambodge sous l'égide de David Levine du Center of Evaluation for Global Action (CEGA) de l'Université de Berkeley. Dans le cas de la seconde, qui constitue notre objet d'étude, le choix du programme a été effectué dans un second temps : une fois posée la volonté de l'AFD et du CEGA de travailler ensemble, en développant cette méthodologie spécifique qu'est la randomisation, le département évaluation s'est tourné vers la division opérationnelle pour savoir quel projet pouvait en être l'objet.

Émerge alors un intérêt tout particulier pour évaluer le projet SKY mis en place par le Gret au Cambodge, pour différentes raisons. D'une part, à cette date, alors que les évaluations randomisées investissent fortement le thème de la microfinance, celui de l'assurance santé dans les pays en développement n'a pas été exploré avec cette méthodologie. C'est donc un sujet porteur à la fois d'un point de vue académique mais aussi stratégique pour l'AFD qui met depuis 2004 l'accent sur le secteur de l'accès aux soins de santé, au sein duquel il existe de nombreux débats sur les approches à privilégier (assurance, transferts d'argent conditionnels, gratuité des soins etc.). D'autre part, le gouvernement Cambodgien réfléchit au développement à court ou moyen terme de sa politique de santé, il apparaît donc particulièrement intéressant d'étudier l'impact réel de ce programme de micro-assurance pour appuyer ou non, auprès des autorités locales, son institutionnalisation au sein d'une politique publique nationale. Enfin, ce programme, jusqu'alors dans une phase pilote, est en passe d'être étendu à de nouvelles zones sur financement de l'AFD, ce qui permet d'évaluer un dispositif stabilisé mais qui s'adresse à des populations totalement vierges, permettant donc de constituer *a priori* des groupes tests et des groupes témoins de manière totalement aléatoire. En outre, le dispositif de SKY utilise déjà un système de loterie promotionnelle pour favoriser l'adhésion de nouveaux membres, qui laisse penser qu'il se prête particulièrement bien à une méthodologie fondée sur la sélection aléatoire des échantillons.

A travers ce choix, l'AFD poursuit donc plusieurs objectifs : l'influence des décideurs politiques dans l'optique de favoriser l'application d'« *evidence-based Policy* » dans la droite ligne des débats sur l'efficacité de l'aide ; l'alimentation de sa propre stratégie de financement dans un secteur en plein développement ; la production de connaissances scientifiques de haut niveau pesant sur le débat international ; et l'acquisition d'expertise sur une méthodologie d'évaluation qui suscite un fort engouement au sein de la communauté internationale, permettant *in fine* de définir une stratégie concernant l'utilisation de cette méthode en interne.

Pour l'AFD, la mise en place d'un partenariat avec le Cambodge pour l'amélioration de la formulation des politiques publiques est un enjeu fort de l'étude. En avril 2006, le département évaluation de l'AFD reprend contact avec l'équipe CEGA pour confirmer leur intérêt mutuel à travailler sur l'évaluation d'un programme de micro-assurance santé au Cambodge. L'agence précise ses objectifs qui sont en effet d'ordre méthodologique, opérationnel, mais aussi partenarial : (1) développer de l'expertise dans le domaine des évaluations à travers le partenariat avec le CEGA, (2) travailler en lien avec des équipes et des autorités cambodgiennes pour

ainsi avoir une influence sur la stratégie du Cambodge en matière de santé, et (3) connaître l'impact de ses interventions pour améliorer ses stratégies sectorielles.⁴

Il ne s'agit donc pas seulement d'influencer le gouvernement Cambodgien dans la formulation de sa politique de Santé, mais bien de contribuer au renforcement des capacités locales à entrer dans ce débat en s'en appropriant les termes et les enjeux.

Le cahier des charges initial de l'AFD est assez large, il décrit les différentes questions d'intérêt pour l'AFD au sujet de la micro-assurance santé : son impact sur la pauvreté et la vulnérabilité des ménages, sur la consommation de soins de santé et d'autres services publics (école, etc.), sur les caractéristiques et le comportement économique des ménages (épargne, employabilité, etc.), et sur la qualité et l'organisation de l'offre de soins. Par ailleurs, l'AFD souhaite que l'impact soit mesuré en fonction de différents scénarii possibles d'intervention : promotion seulement de la micro-assurance ou de celle-ci et à la fois de l'amélioration de la qualité de l'offre de soins, ou seulement de cette dernière quand il n'y a pas d'offre de micro-assurance, l'impact en fonction des différents niveaux de primes demandés, et enfin en fonction de différents niveaux d'information sur les bénéfices de la mutualisation des risques. La question étant de savoir quelles sont les meilleures options pour maximiser l'efficacité de la micro-assurance sur différents types de population (ou pour maximiser l'amélioration de leur santé si la micro-assurance n'est pas concluante).

1.2 Des positionnements divergents très affirmés dès l'origine

En juin 2006, l'équipe CEGA envoie une première proposition à l'AFD, proposition illustrative générale à retravailler avec les acteurs locaux et le commanditaire⁵. Celle-ci est transmise au Gret par le représentant de l'AFD au Cambodge, ce qui constitue le point de départ des discussions, incompréhensions mutuelles et désaccords qui vont marquer le déroulement de l'ensemble de l'étude.

Dès l'origine, l'opérateur est inquiet des effets que pourrait avoir la mise en œuvre d'une grosse évaluation de ce type sur le déroulement des nouvelles phases de SKY. Le document envoyé par CEGA, fondé sur une première réflexion purement théorique, déclenche donc une réaction très vive de la part du Gret. D'une part à cause de son caractère déconnecté de la réalité cambodgienne, et d'autre part des potentielles perturbations que l'étude, à travers le schéma proposé, risque de causer au projet (échanges d'emails du 3 juin 2006).

Cette première proposition générale très théorique du CEGA est prise « au pied de la lettre » par le Gret qui discute chacun des objectifs et des propositions méthodologiques, en premier lieu la taille de l'échantillon (25 000 personnes soit 20% de la population totale de la province d'extension du projet), la pertinence de tester différents systèmes tels que le co-paiement et les différentiels de primes d'affiliation, ainsi que les tests marketing évoqués pour améliorer les taux d'adhésion à SKY. Ces différentes options ont été testées par le Gret depuis plus de quatre ans, donnant lieu à de nombreuses études et conclusions qui ne lui semblent pas être prises

⁴ *"In our understanding, impact evaluation should not result in northern discussion on southern data. A critical condition for us to launch this kind of work is the commitment of Cambodian teams and authorities. It seems possible according to our very preliminary contacts. A clear vision of impact of the micro-insurance system is likely to influence the Cambodian health strategy."* Mails JD: 28 Avril, 2006 12:31 AM JDN to DL

⁵ 2 CEGA RESIPC 030606

en compte dans cette proposition. L'opérateur est également préoccupé par la dimension **éthique**⁶ de certaines options envisagées, telles que la randomisation des centres de santé en fonction de la qualité des soins qui y est proposée, et plus généralement par le fait de maintenir volontairement une partie de la population hors de portée du programme (groupes témoins) pour les seuls besoins de l'étude – ce qui constitue l'une des principales critiques habituellement formulée à l'encontre de cette méthode d'évaluation.

Dès l'origine, le GRET et le CEGA adoptent des positions assez classiques, voire caricaturales, respectivement de l'opérateur et du chercheur. Cependant, différents éléments vont permettre de dépasser cette opposition : d'une part le fait que les principaux intervenants du GRET aient un niveau de compréhension élevé du langage académique – le représentant de l'ONG au Cambodge est par exemple titulaire d'un diplôme de troisième cycle en économie, réalisé dans une université américaine – et une importante culture de la recherche et de la capitalisation sur les projets menés, liée à l'histoire et au mandat de l'organisation, qui fait qu'ils comprennent l'intérêt de la démarche entreprise. D'autre part, le fait que les membres de l'AFD qui suivent ce projet d'évaluation ont une double casquette de bailleurs mais également de chercheurs en économie quantitative spécialisés sur les questions d'évaluation d'impact, tout en ayant la volonté de suivre de très près ce projet et non seulement de le financer et d'en attendre les résultats. Le bailleur va jouer tout au long de l'étude un rôle fondamental de médiateur permettant d'atténuer les tensions entre acteurs. L'AFD va ainsi, dès le départ, et tout au long du projet, insister auprès du CEGA pour que celui-ci prenne en compte les demandes et le point de vue du Gret. Les chercheurs de Berkeley montrent également qu'ils souhaitent jouer le jeu, évoquant à plusieurs reprises leur souhait de dialoguer avec l'équipe du GRET.⁷

Enfin le Gret, conscient de son incapacité à empêcher que cette étude ait lieu, s'engage également très rapidement dans une attitude de compromis : « *puisque'on va sans doute devoir faire avec, autant essayer de calmer le jeu et de trouver des solutions constructives* »⁸.

Ces premiers échanges présagent du rôle déterminant que vont jouer les relations entre les différents acteurs du projet en fonction de leurs intérêts et positionnements qui, eux-mêmes, évoluent au fil du temps. On voit là le poids que peut prendre le point de vue de l'opérateur : dans sa proposition initiale CEGA ne fait que répondre – certes, de manière parfois maladroite et déconnectée de la réalité du pays étudié – aux questions assez larges posées par l'AFD, notamment en ce qui concerne le test des différentes alternatives d'intervention dans le secteur de la santé, pour tirer des conclusions permettant directement d'orienter la politique cambodgienne. Du point de vue de l'opérateur, ces propositions sont inconcevables dans la mesure où le projet SKY sort d'une longue phase pilote ou de nombreux tests ont été effectués, même s'ils n'ont pas nécessairement la validité scientifique des évaluations randomisées. Cette volonté commune du GRET et de l'AFD de limiter au maximum les perturbations opérationnelles de l'évaluation sur le programme, et donc de faire admettre à tous que « *c'est à l'évaluation de*

⁶ Les divergences sur la dimension éthique vont apparaître en filigrane durant tout le processus, d'une part à travers le montant important du financement de l'étude (au total plus d'un million de dollars) qui dépasse largement le budget du programme lui-même, qui est formulé à plusieurs reprises comme un problème éthique par différents acteurs ; d'autre part concernant les questions que CEGA souhaite poser lors de l'enquête – notamment concernant le rapport au risque des affiliés – qui, n'étant pas compréhensibles par les enquêtés, sont contraires à l'éthique du point de vue de Domrei.

⁷ “*At this point, as described in the attached document, it is very important that we at CEGA speak to experts at GRET to find out more about what has already been done to evaluate the SKY program, and about what type of evaluations are realistic and worthwhile for the future*” **Mails CS: 16/08/2006** 12:49 RP to CS

⁸ **Mails CS : 14/07/2006** 04:05 CS to D. Grebert

*s'adapter au projet et pas l'inverse*⁹ » va être déterminante dans la définition et la mise en œuvre de l'étude elle-même.

1.3 La question du partenariat avec les chercheurs « locaux »

Dans ce type d'étude, et cela s'applique plus généralement aux projets de développement, le recours à un partenaire local ayant une bonne connaissance du terrain pour mettre en œuvre le dispositif d'enquête représente une problématique importante. Le choix du partenaire et ses compétences en matière de recherche constituent des points d'interrogations majeurs, notamment lorsqu'on s'intéresse aux processus de production des données statistiques, à leur qualité et à leur validité scientifique. Dans le cas de l'étude d'impact de SKY, paradoxalement, ce n'est pas le manque de qualification ou de rigueur du partenaire qui a posé problème, mais bien au contraire son niveau de compétence élevé qui l'a conduit à vouloir s'impliquer fortement dans la définition des objectifs et de la méthodologie de l'évaluation. Cela est lié d'une part au fait que les directeurs de l'organisme sont occidentaux (Suisse et Australien), ce qui a probablement atténué davantage l'asymétrie de pouvoir entre eux et les chercheurs de Berkeley que s'ils avaient été cambodgiens, et d'autre part à leurs bonnes relations avec les intervenants du Gret qui leur a permis d'unir leurs forces pour peser davantage dans le débat.

Or, les chercheurs de Domrei ont une sensibilité et des trajectoires qui les rapprochent davantage de l'approche compréhensive, dite qualitative, que de l'approche quantitative matérialisée par la démarche RCT adoptée par CEGA (cf.infra). Les conflits qui vont le plus paralyser le processus au fil du temps ne seront pas ceux opposant les chercheurs à l'opérateur, mais bien les chercheurs entre eux, avec un clivage ayant des dimensions à la fois politiques, épistémologiques et méthodologiques lié en partie à la division Nord/Sud.

Celle-ci prend en effet une importance particulière dans ce cas précis car, d'un côté, le but est officiellement pour l'AFD, mais aussi pour les promoteurs des RCT tels que le CGDev – dont D. Levine est un membre actif – de renforcer les capacités des chercheurs des pays du Sud à réaliser des études d'impact dites rigoureuses de projets de développement mis en œuvre dans leurs pays. Cela implique de ne pas considérer le partenaire local comme un simple exécutant producteur de données, que seule l'équipe de recherche « Nord » est en mesure d'analyser, mais bien comme un membre à part entière de l'équipe de recherche, ce qui est le cas de Domrei dans l'étude SKY. Mais d'un autre côté, Domrei est parfois difficilement considéré comme une équipe de chercheurs locaux du fait de l'origine « Nord » de ses directeurs, et surtout, d'un point de vue contractuel, il est sous-traitant de CEGA et n'a donc pas à outrepasser les fonctions que l'équipe de Berkeley lui attribue¹⁰.

Outre les débats sur la méthodologie elle-même, les points de vue des chercheurs, et de l'ensemble des acteurs, vont fortement s'opposer sur la définition même des objectifs de l'étude.

⁹ Mails CS: 14/12/2006 08:56 CS to C Poursat, A Duffau

¹⁰ « ...il faut d'un autre côté bien faire comprendre à Domrei que, en tant que sous-traitant, sa voix en ce qui concerne le design de l'évaluation n'est pas prioritaire. Lorsque la démarche de CEGA fait partie du protocole, il faut la respecter » Propos stagiaire AFD 190609 Point de vue personnel sur la collaboration CEGA Domrei

2. La définition des objets de recherche : le débat impact versus sélection.

Suite aux premiers échanges sur la proposition préliminaire de CEGA, les différents acteurs se rencontrent pour la première fois lors d'une mission d'identification conjointe de l'AFD et de CEGA (qui envoie deux doctorants, mais pas le chercheur principal David Levine) au Cambodge en décembre 2006. Les chercheurs de Berkeley prennent alors conscience de la difficulté de mettre en œuvre une méthodologie randomisée pour évaluer ce programme du fait de la très grande volatilité de ce qui pourrait constituer leurs échantillons (groupes test et contrôle) lié aux mouvements d'adhésion/résiliation très fréquents des assurés. Ils décident de procéder à une sélection aléatoire individuelle et non par villages, dans le programme rural, pour tirer profit du système de loterie promotionnelle déjà mis en œuvre par SKY pour promouvoir le programme. Enfin ils rencontrent les chercheurs de Domrei, bureau d'études cambodgien qui sera chargé de réaliser localement les enquêtes et le suivi de l'étude, pour le compte de CEGA.

Après cette première mission, les chercheurs californiens demandent au Gret de mettre en place des tests sur l'attribution promotionnelle de coupons dans les villages (les *Lucky Draw Coupons*, LDC) selon différentes modalités pour évaluer les taux d'adhésion et de résiliation et ainsi juger de la faisabilité de l'étude à travers cette forme de randomisation.

Suite à cela, CEGA envoie à l'AFD entre janvier et avril 2007 différentes versions de leur proposition de recherche, qui sera finalement approuvée, après commentaires et négociations, en avril 2007, sous réserve de l'approbation du financement de l'extension du programme SKY par la division développement humain de l'AFD, qui n'est alors toujours pas acquise.

La toute première proposition a en fait été écrite pour demander une subvention de recherche au programme Basis de l'USAID qui finance également des études d'impact¹¹. Cela fait apparaître les premières tensions qui tout au long du processus vont se concentrer sur la nature des objectifs de la recherche et surtout sur leur hiérarchisation.

En effet, la proposition Basis est centrée sur l'étude de la sélection adverse, or l'AFD précise que ce document est un bon point de départ mais qu'en aucun cas il ne constitue le projet de recherche qu'ils comptent financer. Ce qui les intéresse en tout premier lieu c'est de savoir si l'assurance santé est un bon moyen pour augmenter les impacts positifs sur la santé et diminuer la vulnérabilité des populations pauvres. Ils sont très intéressés par l'étude de la sélection adverse mais insistent également sur le fait que, pour comprendre totalement la question de la « durabilité » financière d'un système d'assurance, il faut aussi prendre en compte les dimensions liées à l'aléa moral, à la qualité et aux différentes options du service offert. Ils insistent par ailleurs sur l'importance d'une étude qualitative pour comprendre les déterminants de l'adhésion à SKY. Par ailleurs, l'AFD s'interroge, dans l'optique où l'étude ferait l'objet d'un cofinancement AFD-Basis, sur la question de la répartition : qui finance quoi exactement ?

Dans la proposition envoyée à l'AFD quelques semaines plus tard (mi-février)¹² les objectifs sont clairement hiérarchisés : 1. L'impact sur l'accès aux soins et le bien-être économique des affiliés 2. La sélection et les mécanismes d'autosuffisance financière 3. L'aléa moral à travers des enquêtes qualitatives 4. Les effets de l'assurance sur l'offre de soins.

¹¹ 240107 BASIS_proposal

¹² 130307AFD_proposal-1

C'est à partir de là que les discussions entre chercheurs, opérateur et bailleur sur les objectifs de recherches et leur hiérarchisation, au-delà des nombreuses questions de méthodologies soulevées, vont commencer. En effet, l'AFD explique à Berkeley que, de son point de vue, les enjeux de l'étude de la sélection et de l'autosuffisance financière de SKY sont d'améliorer le programme (pouvoir abaisser les primes d'affiliation par exemple) et non de prouver son efficacité (pour décider de poursuivre ou non son financement). De même, l'aléa moral n'est pas un problème dans le cadre du programme puisque le souci actuel réside au contraire dans la sous-consommation de soins dans les centres de santé publics. L'AFD rappelle à CEGA qu'étudier la viabilité financière du système ou travailler sur le *business plan* du programme ne fait pas partie des objectifs de l'étude (échange d'emails du 23 mars 2007).

Le GRET est aussi assez inquiet de l'importance donnée aux thèmes de l'antisélection et de la surconsommation de soins, et insiste sur l'importance, pour comprendre l'impact du programme, de « *tenir compte de l'amélioration du service au niveau des structures de soin* » (échanges d'emails du 26 mars 2007)¹³. D'autant plus qu'il s'agit de son point de vue de données très sensibles, qui, mal interprétées et largement diffusées dans le cadre de publications scientifiques, pourraient injustement condamner le programme SKY.

Suite aux remarques de l'AFD et du Gret, la proposition finalement acceptée en avril 2007 présente des objectifs ainsi hiérarchisés : 1. Impact sur la santé et l'économie des ménages 2. Sélection 3. Changements dans les comportements d'accès aux soins 4. Impact sur l'offre de soins (*providers*).

Mais tout au long de l'étude, les discussions vont se crispier autour d'un débat sur la hiérarchisation des priorités : l'opérateur et le partenaire local chargé de mettre en œuvre l'enquête (Domrei) insistent sur l'étude de l'impact du programme, accusant les chercheurs de Berkeley de ne s'intéresser qu'à la sélection adverse.

Suite au comité de pilotage qui a lieu au lancement de l'étude en février 2008, Domrei fait ainsi part directement à l'AFD de ses préoccupations en la matière, considérant que « [the] overall evaluation design is a poor match for measuring SKY impact but is well suited to studying selection into insurance. »¹⁴

Ils vont même jusqu'à soupçonner l'équipe CEGA - suite notamment à leur demande d'obtenir l'intégralité de la base de données du programme SKY début 2009 - de poursuivre un agenda « secret » en la matière, la doctorante de Berkeley principalement chargée de l'étude par David Levine faisant sa thèse sur le sujet, alors particulièrement porteur dans le champ académique.

*“In situations like this, where one member of a group supposedly working towards a common goal seems to be very out of step with the other members - I suspect that members actually have different goals.”*¹⁵

Cette suspicion se renforce lorsqu'en juin 2009, David Levine veut ajouter 16 nouvelles questions à l'enquête ménages car il veut « savoir si l'assurance volontaire dans les aires rurales du Cambodge (et par extension dans d'autres zones pauvres du monde) peut être financièrement durable »¹⁶.

¹³ Mails CS : 26/03/2007 12:22 C Phily to CS

¹⁴ Mails CS: 26/02/2008 13:05 Domrei to JDN

¹⁵ Mails CS : 05/02/2009 07:44 Domrei to CS

¹⁶ “We would like to know if voluntary insurance in rural Cambodia (and, by extension, in other poor parts of the world) can be financially sustainable. Financial sustainability is challenged by adverse selection because sick people value

Domrei transfère cette information au GRET en ajoutant qu'au téléphone, D. Levine a été beaucoup plus radical dans la formulation de son hypothèse : *"the hypothesis he was testing was that "adverse selection made SKY inappropriate for the rural poor and that other financing options needed to be considered"*.

Le Gret s'empresse alors de contacter l'AFD pour exprimer ses inquiétudes et son désaccord, considérant que CEGA n'a pas compris l'objectif réel de SKY – *« Le but n'est pas de viser certains segments de marché à faible risque mais de couvrir le segment le plus large de la population, même s'il y a un coût supplémentaire que nous sommes prêts à assumer »*.¹⁷

Si l'AFD fait tout pour apaiser les tensions et jouer le rôle de médiateur entre les parties prenantes, le flou qui règne au niveau de la gouvernance globale de l'étude et notamment du fait de ses multiples sources de financement, permet à cette opposition de se maintenir tout au long du dispositif. En effet, quand l'AFD questionne CEGA sur son intérêt prononcé pour cette question de la sélection, les chercheurs répondent qu'elle est traitée dans le cadre de la subvention Basis.

Finalement, quelle que soit la pertinence de cette opposition, que les doutes de certains aient été fondés ou non, l'analyse de ces éléments montre avant tout que ce projet d'évaluation a été mené à son terme malgré les intérêts – nécessairement – divergents des acteurs, mais surtout **en l'absence d'un consensus sur l'objet d'étude lui-même et sur les objectifs finaux de la recherche**, qui transparait clairement dans les débats méthodologiques qui l'ont traversée (cf. infra).

3. L'impact de la recherche sur l'opérationnel, et vice-versa

Suite à la toute première proposition théorique formulée par Berkeley en juin 2006, la principale préoccupation du GRET est de limiter l'impact de cette évaluation sur ses activités opérationnelles. L'opérateur, conscient de l'asymétrie du rapport de force qu'il peut entretenir avec l'AFD, ne s'est jamais opposé jamais ouvertement à ce projet malgré les importantes réserves qu'il peut avoir à son égard. Il adopte une attitude ouverte à la négociation et au compromis, pour mieux faire valoir son point de vue.

Pour limiter au maximum les effets potentiellement négatifs de l'étude sur le programme, sa stratégie consiste premièrement à affirmer et institutionnaliser son rôle dans le pilotage de l'étude et à défendre sa propriété intellectuelle sur les sources des données pour être informé de toutes les publications qui pourraient en être issues ; deuxièmement à tenter de délimiter le plus clairement possible le périmètre de l'évaluation pour limiter sa collaboration aux termes de références établis ; et enfin troisièmement à obtenir des ressources (humaines et financières) pour compenser les effets directs de l'étude sur le dispositif opérationnel.

3.1 Le pilotage de l'étude : quand l'opérateur s'incruste parmi les chercheurs...

Il apparaît clairement dès le début du projet que l'AFD souhaite associer le Gret aux discussions sur le montage et les objectifs de l'étude, non seulement pour assurer sa faisabilité du point de vue opérationnel, mais aussi parce que les deux institutions ont l'habitude de travail-

health insurance the most. Financial sustainability is more likely if (a) there is positive selection or (b) we can help SKY identify targets for marketing that have less adverse selection." Mails JD: 03/06/2009 11:21 CS to JD

¹⁷ Mails JD: 03/06/2009 11:21 CS to JD

ler ensemble et que le bailleur ne souhaite pas perturber le déroulement d'un programme qu'il finance par ailleurs. Le Gret va parallèlement avoir une stratégie assez fine pour imposer son point de vue dans les choix effectués, tant méthodologiques que techniques.

Tout d'abord, la maîtrise des enjeux théoriques de l'évaluation et des contraintes techniques de sa méthodologie, renforcée par le fait que l'équipe du GRET met en place sur le terrain les tests des coupons promotionnels pour le compte de CEGA lui octroie un bon niveau de crédibilité aux yeux des chercheurs de Berkeley, ce qui peut contribuer, en soi, à en faire un acteur important au sein du dispositif. Le GRET en est conscient et cherche à jouer sur cet aspect lors des discussions de fond. Ainsi, lorsqu'à la demande de l'AFD et du CEGA, le GRET envoie ses commentaires sur le projet de recherche en cours de formulation, il indique qu'il souhaite être davantage qu'une simple souris de laboratoire, et argumente de manière détaillée ses inquiétudes, tant sur la faisabilité de la randomisation que sur les possibles perturbations opérationnelles de l'étude sur le programme.

En outre, c'est le GRET qui propose l'équipe de Domrei, avec laquelle il entretient des relations amicales, comme partenaire chargé localement de réaliser le travail d'enquête pour CEGA. Tout en étant parfaitement cohérent et justifié par le niveau et le type de compétences du bureau d'études cambodgien, ce choix va également permettre aux deux organismes de faire souvent cause commune pour défendre leur point de vue contre les propositions ou décisions de CEGA qu'ils n'approuvent pas.

C'est Domrei qui, en avril 2007, envoie une proposition¹⁸ aux chercheurs du CEGA pour préciser et optimiser, au sein du projet final soumis à l'AFD¹⁹, l'implication des différents acteurs, et notamment celle du Gret, dans l'étude d'impact. Le GRET est ainsi considéré comme membre à part entière de l'équipe de recherche²⁰.

Domrei propose par ailleurs la signature d'un *Memorandum of Understanding* (MoU) entre CEGA, Domrei et le GRET, qui spécifie : leur qualité de membre du comité de pilotage de l'évaluation, les rôles et les responsabilités de chacun quant à la conception et la mise en œuvre de l'étude, les règles en matière de propriété intellectuelle, signature des publications, citations etc. La rédaction de ce MoU va être l'objet de nombreuses discussions et l'occasion pour l'AFD de repréciser au Gret - qui n'est toujours pas convaincu par la pertinence de la méthode pour répondre aux questions posées - l'injonction de départ qui est à l'origine de l'évaluation et le fait que le choix de la méthode n'est pas négociable. Le message est clair, le pilotage pluri-acteurs peut orienter et influencer de nombreux aspects de l'étude, mais les limites sont clairement fixées quant aux discussions méthodologiques, l'objectif principal de la démarche, qui est de produire une évaluation randomisée, ayant été fixé en amont. En dehors de cela, tous les commentaires émis par le GRET sont intégrés à la rédaction du document, finalement approuvé le 28 février 2008.

3.2 Des perturbations opérationnelles ?

Finalement, la capacité du GRET à réellement peser sur les orientations de l'étude, sa méthodologie et ses objectifs de recherche est restée très limitée. Les points que le Gret a réussi à

¹⁸ 240407 GRET involvement

¹⁹ 260407 AFD_proposal

²⁰ « *The evaluation team is composed of CEGA (David Levine, Rachel Polimeni, Raj Arunachalam), AFD (Jean David Naudet, Sarah Marniesse, Jocelyne Delarue), DOMREI (Ian Ramage and Gabriel Pictet), GRET (Cedric Salze and Caroline Phily) and RUPP, Nhung Hema* » 260407 AFD_proposal page C.

imposer ont selon nous, paradoxalement, renforcé le caractère restreint des résultats de l'étude, lié aux questions qu'il est possible de poser à travers une méthodologie de type RCT. L'opérateur misait sur un important volet qualitatif pour répondre aux problématiques qu'il jugeait importantes, mais celui-ci a été marginalisé (cf. infra). Par contre, sa stratégie pour réduire au minimum les perturbations opérationnelles liées à l'étude d'impact, qui constituait sa principale crainte au démarrage du projet, a très bien fonctionné.

Tout d'abord, la méthodologie s'est adaptée au schéma du programme en utilisant le système préexistant de loterie promotionnelle – choix qui a justement limité les possibilités de la recherche du fait de cette randomisation individuelle et non par villages²¹. Si ce choix a nécessairement occasionné une charge de travail supplémentaire pour le personnel du programme SKY, notamment en 2007 lors de la mise en œuvre du test des coupons promotionnels, et si globalement, cette étude a constitué une importante débauche de temps et d'énergie de la part des responsables opérationnels sur une période de presque 5 ans au total, de leur propre aveu, ces inconvénients ont été presque entièrement compensés par les apports en termes de moyens qu'a induit l'évaluation (entretien Cédric Salze le 12 octobre 2011).

En effet, non seulement le GRET a obtenu le recrutement au sein de SKY d'une personne dédiée à la saisie des données pour l'évaluation, ainsi que l'achat d'équipement bureautique et informatique pour que celle-ci puisse travailler. Mais surtout, le financement par CEGA des coupons promotionnels de 8 mois gratuits d'affiliation à SKY (au lieu de 1 mois dans le cadre de la procédure habituelle du programme) pour constituer l'échantillon d'étude (cf. infra), a représenté un apport non négligeable de financement et a permis l'accélération de la diffusion du programme dans la province de Takeo. Si les taux d'abandon (*drop out*) des membres affiliés grâce à cette promotion spéciale ont également été importants, ceux-ci ont été partiellement compensés par l'effet d'exemplarité – étudié dans le cadre des monographies de villages réalisées parallèlement à l'enquête ménages – d'un nombre d'affiliation plus élevé que ce que SKY aurait pu susciter sans cette promotion spéciale financée par le projet de recherche²².

4. Un heureux dénouement

Ce qui nous a particulièrement marqué lors de l'étude de ce processus d'évaluation d'impact, c'est le contraste entre le degré de conflictualité et de complexité du dispositif, qui a suscité de vifs échanges et des incompréhensions entre acteurs aux positions largement divergentes, et le degré élevé de satisfaction finale de l'ensemble des participants vis-à-vis des résultats obtenus.

Plusieurs éléments ont été déterminants en la matière. Les deux premiers expliquent que l'étude ait pu être menée à terme malgré des difficultés qui auraient pu être parfois considérées comme insurmontables. Le troisième contribue à ce qu'en plus de ne pas être un échec, cette étude soit même aujourd'hui considérée comme une réussite.

D'abord, l'implication des chercheurs du département évaluation de l'AFD et leur niveau de qualification a permis de jouer un rôle de médiateur efficace, du fait de son respect pour

²¹ Quand le GRET réalise suite au premier comité de pilotage début 2008, que l'étude va se concentrer sur 20% de la population concernée par le programme, et s'en inquiète, David Levine explique clairement que cette limite est liée aux modalités de randomisation individuelle "There are two reasons to focus on the price-sensitive segment : Price-sensitive customers are those ones we are operationally able to study, given individual-level randomization of prices. Unless we can randomly select the villages to offer SKY over many hundreds of villages, there is no alternative to this focus." **Mails CS: 17/04/2008** 11:32 DL to CS

²² Ce qui pose plus généralement la question de la neutralité du dispositif d'étude sur l'objet étudié (cf. infra).

l'ensemble des parties prenantes et de son niveau élevé de compréhension de la méthodologie qui l'a conduite à questionner systématiquement et de manière pertinente les propositions faites par le CEGA et de répondre intelligemment aux inquiétudes du GRET et de Domrei sur les questions de fond. Ensuite le budget de l'étude, considérable, a fait qu'après s'être engagé dans ce processus, tous les acteurs avaient intérêt à ce qu'il aboutisse, pour des questions de crédibilité. Troisièmement, le séminaire de restitution organisé à Phnom Penh les 4 et 5 octobre 2011 pour communiquer les résultats préliminaires de l'étude auprès de l'ensemble des acteurs concernés²³ a joué un rôle important pour la légitimation des résultats et donc de l'étude elle-même.

Soulignons tout d'abord que les résultats montrent un impact significatif du programme sur les dépenses des assurés et sur leur niveau d'endettement, ce qui réjouit tout le monde et en particulier le GRET, qui avait à l'origine conçu ce programme pour réduire la vulnérabilité économique des ménages face aux chocs de santé. Mais ce n'est pas le seul élément qui permet d'expliquer l'enthousiasme de tous après cinq années d'un processus très difficile.

Philippe Lavigne Delville affirme que « le travail de l'évaluateur ne s'arrête pas à la production d'une analyse, il doit la communiquer, et tenter de la faire partager...Ce travail de communication se joue à la fois à l'oral, lors des restitutions [...], et à l'écrit (note de synthèse finale, rapport provisoire, rapport final) »²⁴.

En effet, si lors de la circulation des différentes versions provisoires des rapports rédigés par le CEGA²⁵, la controverse s'est poursuivie, elle ne portait pas particulièrement sur la nature des résultats, mais bien davantage sur leur formulation – le GRET trouvant notamment que la manière de présenter les chiffres sur l'impact tendait à le minimiser.

A l'inverse, pour ses présentations orales lors du séminaire de restitution, David Levine a pris soin de formuler ses résultats de manière absolument consensuelle. D'une part, il a mis l'accent sur les résultats positifs en matière d'impact, en s'efforçant de simplifier au maximum son vocabulaire pour être compris d'un public de non initiés. D'autre part, il s'est attaché à présenter ses conclusions sur la sélection, et plus particulièrement sur la sélection adverse, de manière totalement positive, insistant plusieurs fois sur le fait que, pour un programme à vocation sociale telle que SKY, s'adresser aux gens les plus malades – ceux qui coûtent le plus cher – n'est pas un problème, mais qu'au contraire, cela peut être considéré comme un élément d'efficacité sociale. En outre, pour lui donner un caractère beaucoup moins technique et plus accessible que le contenu de ses rapports de recherches, il a fondé en très grande partie sa présentation sur la sélection sur des éléments tirés des monographies de villages entièrement réalisées par Domrei, ce qui a beaucoup plu aux chercheurs de l'organisme cambodgien, qui y ont vu une forme de reconnaissance de leur travail. Enfin, il s'est rendu extrêmement disponible pour discuter et expliquer ses résultats à toute personne intéressée pendant, et en marge du séminaire, et il a tenu à venir présenter spécialement aux agents de terrain du programme SKY des pistes d'amélioration de leur stratégie marketing dans les locaux du GRET.

²³ Outre les membres de l'équipe de recherche, étaient présents le Gouvernement Cambodgien et notamment les représentants du ministère de la Santé, le personnel de SKY, d'autres bailleurs impliqués dans des programmes de santé au Cambodge tels que la GIZ ou la Banque Mondiale etc.

²⁴ PHLD Evaluation page xx

²⁵ Reflétant l'absence de consensus sur les objectifs de recherche, le CEGA a rédigé deux rapports distincts : l'un sur l'impact, l'autre sur la sélection, comme si ces deux problématiques étaient totalement déconnectées.

Mais le succès « rétrospectif » de l'évaluation ne doit pas occulter les écueils issus des divergences entre acteurs qui ont largement influencé la qualité globale du processus.

La supériorité de l'objectif méthodologique sur tout les autres a considérablement nui à la définition des questions de recherche qui n'ont finalement jamais été stabilisées et clairement formulées. L'une des difficultés rencontrées lors de notre étude a été de comprendre, par exemple, si l'un des objectifs initiaux de l'étude était ou non de mesurer l'impact du programme sur la santé de la population, point sur lequel les différents acteurs rencontrés ne sont pas d'accord, et les documents officiels qui posent les termes de références en la matière ne sont pas suffisamment détaillés apporter une réponse précise à cette question. En fait, l'un des principaux objectifs était bel et bien, initialement, d'étudier l'impact sur la santé, mais celui-ci ne pouvant être étudié à travers la méthode choisie, il n'a jamais été précisé, et a vite été abandonné ou reversé sur des études qualitatives qui n'ont pas reçu l'investissement suffisant pour traiter véritablement cette question. Cela pose la question de la validité d'une démarche scientifique dont le questionnement et la problématique n'est pas déterminé et évolue en fonction de ce que l'on peut mesurer à travers une méthode donnée.

Enfin, les liens entre la recherche et ses implications opérationnelles mis à jour par l'analyse du processus sont pour le moins complexe et ambiguës. Si à un premier niveau, on voit que l'opérateur a su manœuvrer pour « limiter la casse » et même tirer quelques bénéfices de l'étude d'impact, une analyse plus approfondie nous conduit à largement mettre en question la neutralité du dispositif sur le projet étudié. Comment une méthodologie peut prétendre isoler l'impact lié au projet de toutes les autres variables observables et inobservables quand elle transforme à ce point le projet lui-même : en augmentant significativement le taux de subvention des assurés, ce qui accélère considérablement la diffusion du programme et sa couverture territoriale, créant un effet d'exemplarité qui n'aurait pas existé sans l'étude d'impact, mais aussi en influençant le calendrier du projet, ses procédures, l'attitude de son personnel etc. Les enjeux méthodologiques soulevés par cette étude sont l'objet de notre seconde partie.

III. LES ENJEUX METHODOLOGIQUES : RETOUR SUR LES CHOIX EFFECTUES AU REGARD DE LA CAPACITE A LES METTRE EN ŒUVRE. LA RANDOMISATION A L'EPREUVE DE LA REALITE.

Une critique forte adressée aux RCT porte sur l'étroitesse des questions de recherche étudiées : les contraintes techniques de la méthode limitent considérablement le spectre des thèmes de recherche, tant du point de vue des « objets » de recherche que de la manière d'étudier les objets en question (Labrousse 2010 ; Ravallion 2008). Ici en l'occurrence, non seulement l'étude s'est limitée à étudier une maigre facette du projet SKY mais plus encore, et compte tenu des contraintes techniques de l'échantillonnage randomisé, c'est finalement un avatar du projet qui a été évalué (Bernard et al. 2011). Par ailleurs, le protocole d'étude a lui-même été l'objet, comme c'est toujours plus ou moins le cas, d'une forme de « bricolage » mais dont la spécificité ici, à l'instar de nombre de RCT, est d'être passé sous silence (Deaton 2010). Enfin, le recours au « qualitatif », que la plupart des parties prenantes de l'étude appelaient de leurs vœux depuis le départ, n'a pas non plus donné les résultats escomptés, probablement faute d'une réelle volonté – les moyens alloués étaient très maigres – mais aussi du fait de

malentendus quant à la notion même de « qualitatif » et d'une sous-estimation des difficultés épistémologiques que suppose la mise en œuvre de méthodes mixtes.

1. Le poids des contraintes de méthode

« On [...] critique souvent [l'expérimentation] avec la boutade de l'ivrogne qui cherche ses clés sous le lampadaire parce qu'il n'y a que là qu'on y voit quelque chose. Le chercheur en est réduit à évaluer ce qu'il peut évaluer ». Ces propos sont tenus par Esther Duflo dans un ouvrage de vulgarisation sur les RCT et qu'elle tente ensuite d'invalider (Duflo, 2009, p. 51). On ne saurait mieux illustrer le poids des contraintes techniques et l'étude SKY en est une illustration exemplaire : elle montre à quel point le choix méthodologique peut non seulement restreindre considérablement le champ d'étude (effets de court terme, focalisation sur un impact « moyen » et occultation des raisons de l'impact) mais plus encore *déformer* l'objet étudié afin qu'il respecte les canons de la randomisation, et en particulier la présence de deux groupes comparables et d'une taille minimale afin que la comparaison statistique soit suffisamment « puissante ».

Parmi les contraintes techniques, on retrouve les contraintes habituelles des RCT, lesquelles limitent fortement les retombées opérationnelles de ce type d'étude (Ravallion 2008)

- La première limite est celle de la **temporalité**, qui avait d'ailleurs alerté le GRET dès le départ : comment peut-on évaluer un projet sur la santé dans un pas de temps aussi court ? Pour des raisons de coût, les RCT dépassent rarement l'horizon de quelques années. Ici les difficultés rencontrées pour maintenir une participation suffisante ont réduit la période étudiée de 3 ans à 12 mois (cf infra). Cette contrainte temporelle amène souvent à se focaliser sur des résultats intermédiaires (Durand & Nordmann 2011): ici en l'occurrence, c'est principalement l'impact économique qui a été mesuré (réduction de l'endettement et des dépenses de santé). L'impact sur la santé a été évalué, mais du fait de la très faible occurrence de gros chocs de santé – prévisible compte tenu du pas de temps très court – les résultats ne sont statistiquement pas valides. Certes, la question des impacts économiques répond aux préoccupations du projet SKY (le projet a pour objectif premier de lutter contre la vulnérabilité financière des familles) mais limite quand même la portée des résultats, en particulier en matière d'implications en matière de politique publique de santé, alors que c'était précisément un objectif de l'étude. On peut également s'interroger sur la sensibilité des effets à l'épreuve du temps (Labrousse 2010). La micro-assurance est un service nouveau, dont les principes de fonctionnement sont étrangers aux pratiques sociales et culturelles locales. On peut donc supposer que son utilisation évolue au fil du temps et que l'impact après 12 mois est loin d'être stabilisé.
- un second problème est celui des **raisons** de l'impact et de la **chaîne de causalité**. Or comme le souligne Ravallion (2008), une étude d'évaluation n'est utilisable par les acteurs que si elle les éclaire sur les raisons de l'impact. Les RCT évaluent un impact « moyen », mais ne traitent pas des mécanismes de causalité (pourquoi l'impact a eu lieu ?). Elles ont pour objectif d'isoler l'effet d'une action spécifique mais ne peuvent saisir les interactions entre différentes actions (Labrousse 2010 : 13). L'analyse des raisons de l'impact supposerait une analyse à la fois au niveau méso (prise en compte du contexte, de la dimension institutionnelle des actions menées, etc.) et théorique, or ni l'une ni l'autre ne sont dans le champ des RCT. Ici en l'occurrence, on est face à un

projet dont l'impact est en large partie conditionné par un facteur qui a quasiment été évacué de l'analyse, à savoir la qualité des soins²⁶. Malgré les insistances répétées de l'AFD pour que cette question soit traitée à part entière²⁷, elle ne le sera au final que de manière marginale. Lors d'un entretien avec David Levine, celui-ci en explique très clairement les raisons : il reconnaît que cette question est centrale mais qu'elle a été évacuée car « impossible à mesurer » (entretien du 6 octobre 2011)²⁸.

- Pour des raisons de puissance statistique, les RCT ont également du mal à traiter la question de **l'hétérogénéité** (qui sont les gagnants et les perdants ? est-ce que le projet fonctionne mieux dans certains contextes que dans d'autres ?). Qu'il s'agisse de l'amélioration des services ou de réplification à plus large échelle, cette question est évidemment essentielle dans une perspective opérationnelle (Ravallion 2008).
- La non prise en compte des **effets d'entraînements** potentiels limite également les conclusions possibles en matière de réplification à plus large échelle. Ici en l'occurrence, on peut supposer l'existence de deux types d'effets d'entraînements. On a tout d'abord des effets de diffusion, d'apprentissage et d'imitation auprès des populations villageoises, qui peuvent être positifs ou négatifs (c'est d'ailleurs l'une des conclusions de l'étude qualitative qui été menée). On a ensuite des effets sur la qualité des centres de soin, qui peuvent également être positifs (c'est l'hypothèse du projet SKY : les cotisations des assurés sont supposés améliorer la santé financière des centres de soin et donc la qualité des soins) ou négatifs, du fait par exemple d'effets de saturation ou d'engorgement (l'attente est une contrainte majeure des centres publics de soin).

2. La construction sociale de la « rigueur » : un bricolage passé sous silence

Une question plus rarement mentionnée dans les critiques sur la randomisation concerne les défis de mise en œuvre de la méthode elle-même. Comme le souligne Agnès Labrousse la technique des essais cliniques, « relève d'un construit social qui se fabrique de manière située, en médecine comme en économie » (Labrousse 2010, p. XX). Les RCT n'échappent pas au « bricolage » habituel des protocoles de recherche, mais ce n'est pas tant le bricolage qui est en cause que l'absence de réflexivité des chercheurs sur ce construit social et les biais potentiels ainsi induits. Dans un entretien (mais pas dans ses écrits), Esther Duflo reconnaît que l'ajustement aux contraintes de terrain est souvent nécessaire, mais sans guère s'interroger ou commenter les conséquences scientifiques de ces ajustements²⁹. Or dans de nombreux cas les contraintes pratiques et éthiques d'échantillonnage nous emmènent loin du monde idéal de

²⁶ La qualité des soins mais aussi les relations avec les centres de soin et plus en amont les autorités cambodgiennes sont fort variables en fonction des provinces, ce qui explique d'ailleurs que le projet SKY ne soit implanté que dans certaines provinces. Or « *the ability to understand location effects on program performance is key to the implications for scaling up* » (Ravallion 2009, p. 20).

²⁷ L'AFD avait notamment insisté pour que le CEGA recrute un spécialiste de santé publique.

²⁸ Cette étude a finalement pris la forme d'une étude des centres de santé menée sur 38 structures et adressée à leur personnel, résultats d'un compromis mal abouti entre les exigences des différents acteurs et à ce qu'il était possible de faire avec un budget réduit par CEGA à 4000 dollars pour deux vagues d'enquêtes.

²⁹ (Esther Duflo *in* Mayneris, 2009, cité par Labrousse, 2010).

l'expérimentation, remettant en question le principe même du tirage aléatoire, de l'absence de biais de sélection et donc de la supériorité supposée des RCT par rapport à d'autres techniques statistiques : c'est l'une des conclusions d'Angus Deaton dans un article très critique sur les RCT paru dans le *Journal of Economic Literature* (Deaton 2010 : 447)³⁰.

Dans l'étude menée ici, le mode de sélection des participants (et des non participants) soulève deux questions. La première est celle de l'objet étudié : les exigences de construction des échantillons et les compromis auxquels cette construction a donné lieu font qu'au final, c'est un « avatar » du projet SKY qui a été évalué. La seconde est celle de la « pureté » réelle des groupes étudiés.

Comme indiqué plus haut, le GRET avait d'emblée posé comme condition que l'étude ne perturbe pas le fonctionnement quotidien du projet, notamment concernant la sélection des participants et des non participants. Le tirage s'est donc appuyé sur les stratégies marketing habituelles de SKY, en les adaptant toutefois aux besoins de l'étude, et avec un défi majeur, souligné d'emblée par le GRET et l'AFD : maintenir tout au long de l'étude une participation suffisante ou tout au moins compatible avec les exigences économétriques³¹. Le projet SKY, et d'ailleurs la microassurance en général, se heurte en effet depuis le démarrage à une contrainte majeure : la faible adhésion des populations au principe de l'assurance puis des taux très élevés d'abandon. La technique marketing du projet SKY consiste à organiser des réunions villageoises en offrant des coupons gratuits de participation à des familles choisies au hasard. Or le taux de participation avoisine habituellement 15%, ce qui était beaucoup trop bas (ou aurait supposé un échantillon d'une taille gigantesque). La première étape de l'étude, qui a donné lieu à de multiples tests, discussions et concertation avec les différentes parties prenantes, a donc consisté à « forcer » les conditions de participation en testant diverses techniques marketing permettant d'augmenter le taux de participation. La principale mesure a consisté à étendre l'offre de coupons gratuits. Alors que le projet offre habituellement un mois de coupon gratuit, il a été décidé d'offrir deux types de coupons : des coupons d'un mois (ce que le projet fait habituellement) et des coupons de 5 mois, qui ont ensuite été étendus à 8 mois car la participation n'était toujours pas suffisante³². A cela ont été couplées d'autres mesures telles que des incitations aux agents d'assurance et des mesures d'hypersensibilisation auprès des assurés³³. Il a même été question à un moment donné d'évaluer l'impact à 6 mois. Si les pratiques marketing de SKY ont permis d'aboutir rapidement à un compromis sur le mode de sélection des participants et des non participants – « ils faisaient de la randomisation sans le savoir », commente ainsi le responsable de l'équipe (entretien 10 octobre 2011) – il n'en reste pas moins qu'au final, le projet évalué s'éloigne sensiblement des conditions habituelles de sa mise en œuvre. Comme le souligne l'AFD : d'une part, les participants ne sont pas ceux qui

³⁰ Il cite l'une des études RCT les plus emblématiques, celle de Miguel et Kremer à propos des vermifuges et de la scolarisation au Kenya : pour des raisons éthiques les participants n'ont pas été choisis par tirage aléatoire mais par ordre alphabétique, or il n'y a aucune raison de penser que le tirage alphabétique soit équivalent à un tirage aléatoire.

³¹ Plus précisément, d'après l'équipe CEGA une différence de participation de 35% entre le groupe de traitement et le groupe de contrôle était le seuil minimal.

³² Aux villages meetings les coupons ont été présentés comme de réductions de 8 mois gratuits ou d'un mois gratuit. Ainsi, le *High Coupon* est un coupon qui offre 5 mois gratuits pour les 6 premiers mois de l'année et 3 mois gratuits dans les 6 derniers. Le *Low Coupon* offre juste un mois gratuit.

³³ Aux réticences habituelles des villageois se sont rajoutés des éléments conjoncturels (une distribution de coupon était initialement envisagée au moment du nouvel an khmer, période au cours de laquelle rituels sociaux et religieux sont des dépenses prioritaires villageois) et des aléas politiques : à un moment donné un ministre avait envisagé d'offrir de l'assurance gratuite dans toute la province visé par l'étude, ce qui évidemment aurait remis en question l'ensemble du protocole.

participent habituellement, et d'autre part le projet évalué est un projet *subventionné* (Bernard *et al.* 2011) et à cela on pourrait rajouter que le projet a également fait l'objet d'une *hypersensibilisation* auprès des assurés. En d'autres termes, les mesures prises afin de limiter certains biais en produisent finalement de nouveaux, qui ne sont peut-être pas problématiques du point de vue de la validité interne de l'étude, mais qui apparaissent insurmontables concernant la validité externe de l'étude, c'est-à-dire ses usages opérationnels et politiques (Ravallion 2008).

Lorsque l'une d'entre nous a fait cette remarque lors du séminaire de restitution (n'est-ce pas un projet subventionné qui est évalué plutôt que le projet lui-même ?), l'instigateur de l'étude a répondu que c'était effectivement le cas et que c'était une très bonne remarque, mais ce point n'est nullement mentionné dans les différents documents issus de l'étude.

Les différents partenaires de l'étude n'ont pas été entièrement dupes. Le chef du projet SKY, notamment, exprime son scepticisme courant 2008 lorsqu'il prend conscience du problème – il ne comprend plus du tout l'intérêt de l'étude³⁴ –, mais sans que cela modifie quoi que ce soit en termes de protocole. Le CEGA n'avait de toute façon probablement pas le choix, sauf à abandonner la méthode RCT.

D'autres aspects du « bricolage » de l'échantillonnage posent également question : là encore notre propos ne consiste pas à affirmer l'existence de tel ou tel biais, mais à souligner l'absence de réflexivité, condition pourtant essentielle de toute démarche scientifique.

Les participants n'ont pas été tirés au hasard, mais parmi les villageois qui avaient accepté de participer aux réunions villageoises puis au tirage au sort. Or les réunions villageoises opèrent comme un premier filtre de sélection. Le CEGA évoque cette contrainte lors du séminaire de restitution, mais la résous en comparant les caractéristiques socioéconomiques de la population enquêtée avec la population cambodgienne, concluant qu'il n'y a pas de différence significative. Mais qu'en est-il des variables dites « inobservables », contre lesquelles la randomisation est supposée lutter ? On peut supposer que les participants aux réunions villageoises soient plus disponibles, plus curieux et ouverts à l'innovation, plus proches du leader du village, plus insérés socialement, davantage malades, etc. Par ailleurs, certains échanges emails entre CEGA et l'équipe du projet SKY indiquent que certains villageois ont refusé de participer au tirage au sort³⁵.

Parmi les villages choisis, certains étaient vierges de toute intervention en matière d'assurance, alors que d'autres faisaient déjà partie du programme SKY. Il semblerait que les villages non vierges aient été inclus dans un second temps, précisément afin d'augmenter le taux de participation. Mais ceci n'est pas mentionné dans les résultats, et les effets potentiels d'entraînement sont passés sous silence.

On peut également s'interroger sur un éventuel marché secondaire des coupons. D'après une discussion avec le chef de projet SKY, de multiples précautions avaient été prises et théoriquement rendaient impossible ce scénario (de nombreux emails ont été échangés sur ce sujet, et l'équipe CEGA était bien consciente de ce risque). Or plusieurs éléments indiquent que la question mérite d'être posée, et d'ailleurs plusieurs échanges emails entre l'AFD et le CEGA portent sur ce sujet : on observe des décalages entre la base de données de l'étude et celle du projet, comme si certains participants sélectionnés n'avaient en fait pas adhéré au projet ; on

³⁴ Mails CS: 16/04/2008 20:13 CS to Milton Fong

³⁵ Mails CS du 31/01/2007 RP et C Phily

observe que le profil des assurés ayant bénéficié de « coupons élevés » est significativement différent des assurés ayant bénéficié de « low coupon » (voir les échanges entre l'AFD et CEGA sur cette question). Interrogée sur cette question, l'équipe cambodgienne Domrei a admis que malgré toutes les précautions prises, cette hypothèse était probable : l'équipe a procédé à des vérifications téléphoniques qui ont confirmé que certains n'étaient pas ceux qui avaient été tirés au sort (Entretien 12 octobre 2011, Phnom Penh).

La méthode consiste à évaluer les personnes avant et après l'adhésion au programme, ce qui suppose une première vague d'enquête dès l'obtention des coupons et une seconde vague 12 mois plus tard. Cette contrainte a des conséquences logistiques fortes, puisqu'il faut être en mesure d'enquêter *immédiatement* les bénéficiaires des coupons. En pratique cette condition n'a pas été entièrement respectée : suite à des retards de paiement, chroniques semble-t-il depuis le démarrage du projet, Domrei n'a pas été en mesure de procéder aux enquêtes immédiatement pour une partie de la population ciblée : la première vague d'enquête a été réalisée plusieurs mois après le démarrage, période au cours de laquelle les assurés ont pu bénéficier de soins gratuits. Au final, pour une partie de la population étudiée, la comparaison ne porte pas sur « avec » et « sans » assurance, mais sur différentes durées de statut d'assuré. Là encore, ce point est éludé des résultats de l'étude.

Cette absence de réflexivité – ou du moins le fait qu'elle soit occultée des résultats – pose sérieusement question puisqu'on voit à quel point le protocole de recherche déforme l'objet étudié. On est évidemment très loin de la méthode ethnographique, pour qui la réflexivité est partie intégrante de l'interprétation : l'ethnographe se considère comme un « catalyseur » des relations sociales pré-existantes et ses observations n'ont de sens et ne peuvent être interprétées qu'à travers une analyse réflexive quant à la position du chercheur dans la scène observée. A chaque approche ses méthodes, arguera-t-on, mais on est également très loin des essais cliniques médicaux et de leur fameuse règle du « double aveugle », visant précisément à éliminer la subjectivité des parties prenantes de l'expérimentation, qu'il s'agisse des patients ou des intervenants : ni les uns ni les autres ne doivent savoir la nature du traitement.

3. Le recours au « qualitatif » ? Une tentative avortée

Le principe d'une RCT n'a jamais été remis en question : pour le bailleur (AFD), l'objectif de l'étude visait précisément à tester ce type de méthode. En revanche il a rapidement été question de compléter la randomisation avec des analyses qualitatives. Au cours de toute la phase de construction du protocole de recherche, la nécessité de rajouter puis de renforcer le volet « qualitatif » fait l'objet de commentaires répétés de la part des différentes parties prenantes de l'étude. Le GRET, en particulier, attire régulièrement l'attention sur le fait que le protocole de recherche reste très flou sur la dimension « qualitative ». Au final qu'a-t-il été produit à partir d'une méthode « qualitative » et que peut-on en conclure sur l'articulation entre ces deux types de méthode ?

On observe tout d'abord que l'analyse « qualitative » a été réduite à une portion congrue, à en juger déjà par la répartition des financements – dont moins de 10% ont été consacrés aux enquêtes qualitatives. Elle a d'abord été mobilisée pour contextualiser l'enquête et l'adapter aux

spécificités locales. On ne peut toutefois parler « d'analyse qualitative », mais simplement de mobilisation d'informations contextuelles³⁶.

Concernant l'analyse proprement dite, elle a été réduite à des monographies de villages, discutées de longue date mais menées début 2009, au moment de la seconde vague d'enquête par questionnaire. Elles ont été menées dans d'autres villages que ceux de l'enquête ménages, où le programme SKY existait depuis longtemps afin d'avoir le recul nécessaire pour comprendre les dynamiques d'affiliation. Leur objectif principal était d'étudier les processus sociaux d'adhésion et de rejet de l'assurance (rôle des liens de parenté et de voisinage, du leadership etc.). D'après certains échanges email entre Domrei et l'équipe SKY, il était question au départ que ces enquêtes soient menées en amont et servent à alimenter le design du questionnaire. En pratique toutefois, et probablement pour des raisons budgétaires, elles seront menées plus tard et donc en parallèle de l'enquête par questionnaire. L'étude des centres de soin et du rôle du projet SKY sur l'amélioration de la qualité des soins, auquel l'AFD semblait tenir, devait au départ être menée de manière qualitative (mais sans que le terme « qualitatif » ait réellement été explicité) a progressivement été réduite à quelques entretiens menés auprès du personnel (cf. supra).

Au vu du coût total de l'étude, on ne peut que regretter que la partie « qualitative » n'ait pas été plus conséquente. Mais il aurait certainement fallu considérer autrement le « qualitatif » pour qu'il puisse réellement avoir sa place.

La question de l'articulation entre « qualitatif » et « quantitatif » prête très souvent à confusion du fait des malentendus quant à l'usage même des termes (Winter 1983) et il semblerait que l'étude SKY n'ait pas échappé à cette confusion. A la lecture des documents relatifs à l'étude et des échanges emails, il semblerait que le terme « qualitatif » ait tantôt été utilisé dans une perspective de contextualisation (adapter les questionnaires aux réalités locales et limiter les biais ethnocentriques de l'équipe CEGA), tantôt dans une perspective d'élaboration d'hypothèses qui seraient ensuite testées par l'enquête statistique (cf échanges Domrei/GRET), tantôt dans une perspective d'analyse et de production de connaissance à part entière (cf attentes de l'AFD concernant les interactions entre le projet SKY et l'amélioration de la qualité des soins). Le terme « qualitatif » prête à confusion car il laisse entendre qu'il y aurait des méthodes basées sur des variables numériques et la quantification et d'autres ayant recours à « d'autres » outils. Or les méthodes qualifiées de qualitatives comme l'ethnographie et certaines branches de la sociologie ont très largement recours au comptage, à l'analyse quantitative lexicale de corpus de textes, etc. Inversement de nombreuses enquêtes statistiques introduisent des variables qualitatives dans leurs questionnaires. Il existe en revanche un clivage net de nature épistémologique entre les approches que l'on pourrait qualifier de statistique (le « quantitatif ») et les approches que l'on pourrait qualifier d'interprétative et de compréhensive (le « qualitatif »). Les premières procèdent par extrapolation en faisant l'hypothèse que la réalité est décomposable en éléments semblables, donc relativement simples et identi-

³⁶ Consultation du personnel SKY (direction du projet et agents d'assurances) pour définir les hypothèses et les variables à introduire dans le questionnaire

Mentionner que cette étape, même s'il ne s'agit pas d'analyse qualitative a proprement parler, a mobilisé une certaine énergie – critiques fortes au départ sur le décalage entre les hypothèses posées et les spécificités cambodgienne ; ensuite efforts répétés de Domrei pour que le GRET soit davantage impliqué ; en même temps GRET pas toujours très réactif ? peu enclin à fournir certains documents qui auraient peut-être pu permettre à CEGA d'élaborer des hypothèses plus adaptées ??

fiables, et le rôle de l'analyse est de tester des hypothèses de similitude ou de différenciation de collection d'objets, d'individus ou d'événements. Et seule la réduction de la complexité et de la variété du réel autorise la mise en évidence de similitude et d'homogénéité. Les secondes visent à mettre à jour des modes de fonctionnement et de reproduction d'entités sociales complexes, les interactions multiples, dynamiques, contradictoires entre différentes entités, tout en accordant une attention centrale au point de vue des acteurs (Winter 1983).

L'opposition entre ces deux types d'approches renvoie à des « systèmes de différences » qui opposent deux « constellations » de pratiques réelles ou supposées dont le clivage est souvent réduit de manière superficielle à l'usage d'outils distincts (Couty 1983). Même si ce clivage est souvent simplifié à l'extrême et caricaturé, il n'en reste pas moins qu'il existe et qu'il dépasse de très loin les considérations techniques. Comme le souligne Appadurai (1989, p. 276), bien au-delà des débats relatifs à la taille d'échantillons, aux marges d'erreurs, etc., le clivage est de nature épistémologique. Il renvoie à des différences concernant la nature des objets étudiés, les postulats qui s'y rattachent, au statut que l'on accorde à la production de connaissance, etc. (Kanbur et Shaffer 2007). En introduction d'un numéro spécial de la revue *World Development* consacré aux approches mixtes, Ravi Kanbur et Paul Shaffer (2007) constatent qu'il existe aujourd'hui un certain consensus quant à la nécessité des approches mixtes, mais que la question de la mise en pratique pose des défis multiples et souvent négligés. En introduction d'un ouvrage consacré au dialogue entre économistes et anthropologues, Pranab Bardhan et Isha Ray (1989) insistent sur le respect mutuel comme condition essentielle d'un dialogue constructif. Si elle peut sembler évidente *a priori*, elle ne va pas nécessairement de soi puisqu'elle suppose au préalable un dialogue permettant, par rapport au sujet étudié, de clarifier les positionnements respectifs et d'identifier clairement les points de divergence³⁷.

Ici en l'occurrence, le décalage épistémologique ont rendu le dialogue extrêmement difficile, et le fait que le terme « qualitatif » n'ait jamais été réellement explicité a probablement contribué à entretenir des malentendus. Chaque partie prenante de l'étude (CEGA d'un côté, Domrei/SKY de l'autre) n'a jamais réellement été convaincue de la légitimité de l'approche préconisée par l'autre partie. Domrei/SKY reconnaissent volontiers la nécessité du principe des groupes de contrôle, mais ont toujours été et restent sceptiques sur le principe de la randomisation. De son côté l'équipe CEGA reconnaît volontiers le rôle déterminant des équipes locales (Domrei et SKY) dans le choix des variables d'impact et l'adaptation des questionnaires – le « qualitatif » est considéré comme une étape centrale de contextualisation de l'étude. En revanche que le « qualitatif » soit une méthode possible de production de connaissance semble beaucoup plus problématique, la démonstration passant nécessairement par l'inférence statistique.

L'appréciation finalement positive des responsables de SKY sur les faibles perturbations opérationnelles, exprimée *a posteriori* – qui n'est pas toujours évidente à la lecture des échanges d'e-mails tout au long du processus – est intimement liée aux résultats de l'étude et à la manière dont ceux-ci ont été restitués lors d'un séminaire à Phnom Penh début octobre 2011.

³⁷ L'ouvrage en question portait sur les biens communs, et trois thèmes avaient été identifiés comme points d'achoppement : autonomie *versus* encastrement, résultats *versus* processus, parcimonie *versus* complexité.

IV. CONCLUSION : LES CONTRADICTIONS ENTRE LA DEMARCHE ET LE DISCOURS QUI LA SOUTIENT

Reconstituant les différentes étapes de formulation et d'exécution de ce projet d'étude d'impact, plusieurs éléments nous semblent émerger. D'une part, il apparaît clairement que l'injonction de départ sur la méthode à employer a déterminé tout le reste du processus. Non seulement à cause du fait que la randomisation elle-même restreint le champ des questions de recherches qui peuvent être posées. Mais aussi, car, du point de vue des acteurs, le choix en amont de la méthode à employer a largement restreint le périmètre de ce qui était négociable et des décisions qu'il était possible de prendre sur une base collective. Cet élément renvoie au type d'évaluation que l'on souhaitait mettre en place : si le fait d'inclure l'opérateur au sein du comité de pilotage et de l'équipe de recherche laisse penser que ce projet constitue une évaluation de type pluraliste, ou au moins coproduite, l'étendue réelle de ce qui est négocié contredit cette analyse.

« Ce type d'évaluations préconise de réunir un petit nombre d'acteurs clés du programme pour réaliser la quasi totalité des composantes de l'évaluation [définition du contenu du projet ; validation de la méthode utilisée ; supervision des travaux, collecte des données ; interprétation ; recommandations] en interaction avec l'équipe de chercheurs. Il considère que pour maximiser l'utilisation de l'évaluation, il est nécessaire d'y associer dès l'amont les acteurs ayant le pouvoir d'agir sur le programme sur la base des résultats de l'étude, et de les impliquer dans toutes les phases du processus. » (Baron et Monier 2003, p.3)³⁸

Sortir d'une logique d'audit et aller vers un système de co-apprentissage, impose de discuter collectivement, avec toutes les parties prenantes, de ce que l'on veut évaluer, pour quelles raisons et avec quels outils. Cela permet, dans un contexte où les intérêts des différents acteurs sont nécessairement distincts, voire divergents, de parvenir à des résultats qui soit légitimes et donc appropriés et réellement utilisés pour réorienter l'action – ce qui est l'un des objectifs annoncé plus ou moins explicitement par les promoteurs de la randomisation – sans pour autant nuire à leur validité scientifique.

Mais l'on touche là à l'une des contradictions inhérentes au discours porté par les chercheurs promouvant les évaluations de type RCT : celui-ci nous explique d'un côté qu'il faut absolument que les évaluations soient utiles à l'amélioration des politiques de développement et qu'il faut donc faire en sorte que les bailleurs – et éventuellement les autorités des pays concernés et les différents opérateurs de développement – s'en approprient les résultats. Cela ne peut passer que par une forte légitimité du processus et de ses résultats puisque *« comme toute institution, l'évaluation dispose d'un espace de validité largement dépendant de deux dimensions : sa capacité à engendrer des savoirs pertinents pour l'action et à susciter une forte acceptabilité »* des ses prescriptions, c'est à dire à construire, ou du moins favoriser une *légitimité politique de ses constats et préconisations* ». Le CEGA s'est fortement employé à assurer l'acceptabilité des résultats de l'étude à travers la formulation qu'il en a faite lors du séminaire de restitution, et même à renforcer la légitimité du processus lui-même en acceptant d'intégrer le GRET, Domrei et les autorités cambodgiennes au comité de pilotage de l'étude. Mais cette démarche n'a pas été jusqu'à une véritable « co-production » puisque la validation du contenu du projet et la définition de la méthode à employer n'ont jamais été ouvertes à la négociation. Cela n'a rien de très étonnant quand on réfléchit au fait que les approches plura-

38

listes en matière d'évaluation reposent en partie sur des principes constructivistes alors que les *randomistas*³⁹ prônent, d'un autre côté, un discours fondamentalement positiviste.

En effet, au-delà du débat épistémologique, il convient de souligner que les résultats produits par l'étude n'ont qu'une portée très limitée sur l'orientation de la politique de santé Cambodienne appelée de ses vœux par l'AFD à l'origine du projet. D'une part, parce que ces résultats ne permettent pas réellement la formulation de recommandations générales, d'autre part car les chercheurs et le bailleur n'ont jamais réussi à impliquer les autorités publiques locales dans le processus. Le périmètre de notre étude ne nous a pas permis d'analyser les causes et les ressorts du manque d'intérêt de ces dernières, mais, de notre point de vue, il n'a pas été favorisé par l'opacité des objectifs de recherches et la complexité des débats méthodologiques qui ont animé l'ensemble du processus.

Les promoteurs des RCT reconnaissent eux-mêmes que la randomisation n'est pas la panacée et que la méthode d'évaluation doit être adaptée à la question posée, et souvent appliquée conjointement à d'autres types de méthodes⁴⁰. Or l'équipe de chercheurs a poursuivi des objectifs académiques de reconnaissance plus qu'il n'a cherché à formuler des recommandations utiles à l'action⁴¹. La focalisation sur la question de la sélection ne correspondait absolument pas aux attentes du débat local sur la politique de santé. Certes on ne peut étudier l'impact sans étudier sérieusement la sélection, mais l'emphase quasi-exclusive mise par les chercheurs sur ce sujet et la négligence à l'égard des autres types de méthodes qui auraient pu utilement compléter cette démarche pour élargir le questionnement à des problématiques « utiles » aux décideurs, telles que l'impact sur la santé et sur la qualité de l'offre, fait qu'on se retrouve avec une étude qui n'apporte pas grand chose aux acteurs politiques locaux, ni même aux bailleurs en terme d'orientation de leur action pour l'avenir.

Au sein de l'argumentaire développé par le 3IE pour répondre aux critiques des RCT, l'on trouve une section permettant de justifier le coût de ces évaluations, qui doit être mesuré à l'aune des économies réalisées en programmes inefficaces et des gains qu'un gouvernement peut tirer du fait de répliquer à grand échelle une politique qui fonctionne, sur le principe des *evidence-based policy*. Au-delà de cet argument, qui ne se vérifie pas toujours dans la réalité, le montant élevé de ces études pourrait contribuer à entretenir leur suprématie, peu des acteurs y ayant été impliqué, et notamment ceux qui les ont financés, n'étant disposés à critiquer un dispositif dans lequel ils ont investi des sommes importantes, notamment au regard de la raréfaction globale des ressources consacrées au développement.

Finalement cette évaluation est aujourd'hui considérée, au moins officiellement, comme un succès en dépit du champ assez restreint des réponses qu'il apporte, notamment au regard des objectifs qui étaient fixés en matière d'orientation des politiques publiques, et malgré les importantes questions qu'il soulève du point de vue de la démarche scientifique sur laquelle il s'est fondé. C'est certainement ce qui fait toute la force d'une approche portée et promue de

³⁹ Ref.

⁴⁰ *when will we ever learn* page 79

⁴¹ Cela peut également être lié au fait que les évaluations d'impacts cherchent, comme leur nom l'indique, à évaluer l'impact d'un projet ou d'une politique selon une logique binaire « ça marche/ça ne marche pas », ce qui appelle nécessairement des recommandations du type « on continue/on arrête de financer ». La compréhension des chaînes de causalité qui permettraient de réfléchir sur les processus qui mènent ou non à l'impact, et donc d'améliorer les dispositifs d'action ne fait pas partie de leurs objectifs premiers.

manière extrêmement volontariste et qui sous le sceau d'une scientificité irréprochable véhicule avant tout un discours idéologique tendant à la dépolitisation de l'action publique.

Bibliographie

Bardhan P. et Isha Ray (eds) *Dialogue between economists and anthropologists*, New-Delhi : Oxford University Press, 1989.

BARON E. ERIC MONNIER « Une approche pluraliste et participative : Coproduire l'évaluation avec la société civile. » in *Informations sociales* n° 110, Septembre 2003.

BEDECARRATS F., « Mise en perspective de 154 recherches sur l'impact de la microfinance », *Séminaire sur les études d'impact*, Nogent-sur-Marne, CERISE/GRET, 14 octobre 2010, www.lamicrofinance.org/files/27398_file_Bedecarrats_CERISE_Introduction.pdf

BERNARD T. NAUDET J.-PH. DELARUE J. On 'nailing' what works through impact evaluations. Lessons from experience at AFD, Document non publié, 2011.

BOUQUET E., WAMPFLER B., RALISON E. « Rigueur scientifique et pertinence opérationnelle des études d'impact en microfinance : une alliance à construire. Enseignements d'une étude en partenariat à Madagascar » *Revue Tiers Monde* (197), 2009 pp. 91-108.

CENTER FOR GLOBAL DEVELOPMENT, When will we ever learn? Improving lives through impact evaluations, Washington DC: Center for Global Development, 2006.

KANBUR et SHAFFER 2007.

LABROUSSE Agnès, « Nouvelle économie du développement et essais cliniques randomisés : une mise en perspective d'un outil de preuve et de gouvernement », *Revue de la régulation*, 7(2), Juin 2010, pp. 2-32.

RAVALLION M. Evaluation in the practice of development, *Policy Research Working Paper 4547*, Washington: World Bank, 2008.