

Les évaluations d'impact dans le domaine du développement

État des lieux et nouveaux enjeux

Stéphanie PAMIES-SUMNER

Département de la Recherche, AFD

Les évaluations d'impact dans le domaine du développement

État des lieux et nouveaux enjeux

AUTEUR

Stéphanie PAMIES-SUMNER*

Département de la Recherche, AFD

* Au moment de la rédaction de ce texte, Stéphanie PAMIES-SUMNER travaillait au sein de la division de l'Évaluation.

À Savoir

Créée en 2010 par le département de la Recherche de l'AFD, la collection À Savoir rassemble des revues de littérature ou des états des connaissances sur une question présentant un intérêt opérationnel.

Alimentés par les travaux de recherche et les retours d'expériences des chercheurs et opérateurs de terrain de l'AFD et de ses partenaires, les ouvrages de cette collection sont conçus comme des outils de travail. Ils sont destinés à un public de professionnels, spécialistes du thème ou de la zone concernés.

Retrouvez toutes nos publications sur <http://recherche.afd.fr>

Précédentes publications de la collection (voir page 101).

Cet ouvrage est téléchargeable gratuitement sur www.afd.fr/A-Savoir

[Avertissement]

Les analyses et conclusions de ce document sont formulées sous la responsabilité de son auteur. Elles ne reflètent pas nécessairement le point de vue de l'AFD ou de ses institutions partenaires.

Directeur de la publication :

Anne PAUGAM

Directeur de la rédaction :

Alain HENRY



Conception et réalisation : Ferrari / Corporate – Tél. : 01 42 96 05 50 – J. Rouy / Coquelicot
Imprimée en France par : STIN

Remerciements

Mes remerciements vont à Laurent Fontaine, Bertrand Savoye, Nicolas Meisel et Florent Bédécarrats pour leur relecture et leurs conseils précieux tout au long de ce travail. Ils s'adressent également à l'ensemble des membres du Comité des évaluations de l'Agence Française de Développement (AFD), auxquels ce travail a été présenté en octobre 2013. Enfin, je remercie l'ensemble des personnes, parmi les bailleurs et les centres de recherche, contactées dans le cadre de cette étude, pour leur gentillesse et leur disponibilité.

Sommaire

1. Introduction : un essor marqué des études d'impact dans le domaine du développement	7
1.1. Les évaluations d'impact : une réponse aux besoins des bailleurs ?	9
1.2. Les évaluations d'impact : le « <i>big push</i> » des chercheurs	11
1.3. La mise en place de structures et de financements inédits en faveur des évaluations d'impact	17
1.4. Les évaluations (d'impact) au-delà de l'aide	19
2. <i>What have we learnt?</i> Points de consensus atteints au niveau international	21
2.1. « <i>Size matters</i> » ?	21
2.2. Du « <i>gold standard</i> » aux « <i>good standards</i> »	29
2.3. Une clarification des attentes et des objectifs fixés aux évaluations d'impact	44
3. Nouveaux enjeux posés par les évaluations d'impact	63
3.1. Les évaluations d'impact sont-elles compatibles avec la Déclaration de Paris ?	63
3.2. Comment mieux articuler production de données pour les évaluations d'impact et renforcement des systèmes statistiques nationaux ?	66
3.3. Les évaluations d'impact : un bien public ou un bien de club ?	80
3.4. Quelles perspectives de développement des évaluations d'impact dans un contexte international changeant ?	82
Annexes	85
1. Sources utilisées	85
2. Liste des institutions interviewées	86
3. Quelques prérequis relatifs à la faisabilité et à l'efficacité d'une expérimentation aléatoire contrôlée	87
Liste des sigles et abréviations	89
Bibliographie	93

Liste des encadrés

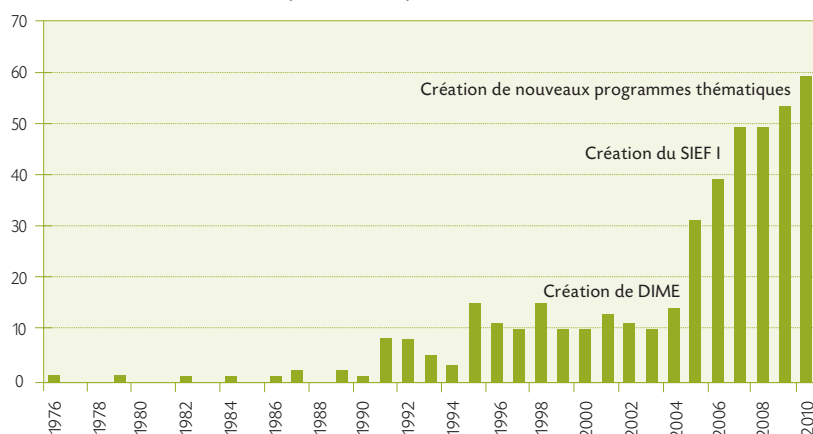
- Encadré 1** Évaluations d'impact : de quoi parle-t-on ?
- Encadré 2** Que peut-on apprendre des évaluations « standards » de projet des bailleurs ? L'exemple de l'AFD
- Encadré 3** Les méthodes promues d'évaluation des politiques publiques : un choix politique ?
- Encadré 4** L'économie des évaluations d'impact : éléments chiffrés à partir de l'expérience de quelques bailleurs
- Encadré 5** Focus sur la méthode expérimentale ou RCT
- Encadré 6** Évaluations d'impact : vers une plus grande mixité de méthodes
- Encadré 7** Évaluations d'impact *versus* autres types d'évaluation : un – désormais – faux débat ? Le cas de la coopération néerlandaise
- Encadré 8** Études sommatives *versus* études formatives : l'expérience de l'AFD, du MCC et de la JICA
- Encadré 9** Quelques exemples d'évaluations d'impact utilisées par des gouvernements des pays partenaires
- Encadré 10** Quelques exemples de résultats novateurs des évaluations d'impact
- Encadré 11** Focus sur la stratégie de DIME et exemples de résultats très opérationnels utiles
- Encadré 12** Comparabilité d'une enquête *ad hoc* avec une enquête nationale : l'exemple d'une évaluation d'impact au Maroc

1. Introduction : un essor marqué des études d'impact dans le domaine du développement

Une progression remarquable du nombre d'« évaluations d'impact » dans le secteur du développement est constatée depuis le début des années 2000. Ainsi, de quelques dizaines à la fin des années 1990, on serait passé à plus de 800 (achevées ou en cours) en 2011, selon l'*International Initiative for Impact Evaluations* (3IE) (White, 2011), et probablement à environ un millier actuellement (Ruben, 2012), bien que leur décompte soit un exercice difficile (cf. encadré 1). Deux types de facteurs ont concouru à cette expansion : un choc de « demande » à la fois diffus et hétérogène du côté des pays donateurs (cf. section 1-1) et un choc « d'offre » marqué du côté des chercheurs (cf. section 1-2).

Graphique 1 Les évaluations d'impact dans le domaine du développement : une progression exponentielle

Évaluations d'impact à la Banque mondiale (année de lancement)



Source : Banque mondiale.

Encadré 1 Évaluations d'impact : de quoi parle-t-on ?

Définition générale

Les évaluations d'impact, au sens strict du terme, sont des évaluations qui visent à *mesurer* les effets *strictement attribuables* à une intervention. Pour ce faire, elles se basent sur la notion de « contrefactuel » : il s'agit ainsi de comparer la situation d'un groupe de bénéficiaires (parfois appelé groupe de traitement en référence au vocabulaire des sciences médicales dont ces méthodes sont issues) d'une intervention à celle d'un groupe dit de contrôle, le plus comparable possible *ex ante* au groupe de bénéficiaires. Différentes méthodes d'évaluation d'impact existent (voir encadré 6), de la méthode expérimentale par assignation aléatoire contrôlée (*Randomized Controlled Trial* [RCT]), en théorie la plus robuste pour éliminer le biais de sélection (différence entre la situation des deux groupes non liée à l'intervention, mais liée à une différence autre, par exemple, si le groupe de bénéficiaires est *ex ante* plus pauvre ou plus malade que le groupe de comparaison), à des méthodes dites quasi-expérimentales, voire à des approches plus qualitatives.

À noter que le terme d'impact a ici un sens légèrement différent de celui défini par le CAD (Comité d'aide au développement) dans son glossaire de l'évaluation. Les « impacts » mesurés dans les évaluations d'impact sont ainsi des effets *causals* d'une intervention, qui peuvent se manifester à court, moyen ou long terme (mais sont le plus souvent mesurés à court terme). Dans la définition du CAD, en revanche, la notion de causalité est moins centrale, tandis que les impacts sont clairement définis comme des effets à *long terme*. Dans les deux acceptions, ces effets peuvent être mesurés/apprécies non seulement au niveau des individus, mais aussi de structures intermédiaires (écoles, centres de santé, entreprises, etc.).

Combien y-a-t-il d'évaluations d'impact dans le domaine du développement ?

Connaître le nombre exact d'évaluations d'impact dans le domaine du développement se révèle un exercice difficile, dans la mesure où il n'existe actuellement pas de base unifiée recensant l'ensemble des évaluations d'impact dans le domaine du développement. Ceci n'est pas surprenant puisque les études fondées sur ces techniques, à la frontière entre évaluation de projet et recherche, ne sont pas nécessairement répertoriées comme des évaluations d'impact en tant que telles dans les bases de données (d'autant plus dans des pays comme la France, où l'évaluation n'est pas considérée comme une discipline académique, à l'opposé de certains pays comme les États-Unis) ^[1].

...

[1] Le *Journal of Economic Literature* (JEL) prévoit néanmoins deux codes qui peuvent se prêter à ces travaux : D04 (*Microeconomics/Microeconomic policy : formulation ; implementation ; evaluation*) ; H43 (*Publicly provided goods/Project evaluation ; Social discount rate*).

...

En outre, même lorsque l'on s'intéresse spécifiquement aux évaluations d'impact produites par les bailleurs (multilatéraux ou bilatéraux), on se heurte également à des difficultés de comptabilité (notamment parce que plusieurs services différents peuvent être à l'origine de ces études – service de recherche économique, service d'évaluation, services opérationnels – et que la définition de ce qu'est une évaluation d'impact peut différer selon les personnes interrogées). Certaines initiatives pour améliorer le recensement des évaluations d'impact ont toutefois été récemment lancées (par le 3IE ; par l'*American Economic Association* pour les expérimentations aléatoires contrôlées en économie).

1.1. Les évaluations d'impact : une réponse aux besoins des bailleurs ?

À l'origine des évaluations d'impact, une demande diffuse des bailleurs vient d'abord se cristalliser sous l'effet de deux facteurs principaux :

- En premier lieu, avec l'adoption des objectifs du Millénaire pour le développement (OMD) par l'Assemblée générale des Nations unies en 2000, la communauté internationale s'engage pour la première fois sur un objectif commun et prioritaire de réduction de la pauvreté. Les 189 pays signataires marquent notamment leur volonté de réduire de moitié l'extrême pauvreté d'ici 2015, dépassant un seul engagement en termes de flux d'aide (comme cela était le cas par le passé) ^[2]. Les OMD constituent également une rupture avec les objectifs passés, en ce sens qu'ils ne se limitent pas aux seules grandeurs macro-économiques (comme le revenu par habitant) ; ils donnent de la pauvreté une définition multidimensionnelle, qui vise à mieux apprécier les conditions de vie effectives des ménages : accès à l'éducation, à la santé, à l'eau potable ; égalité des genres ; etc.
- Dans le même temps, l'engagement international renouvelé en faveur de l'aide au développement, avec l'adoption des OMD, s'accompagne d'une volonté forte d'améliorer « l'efficacité de l'aide », illustrée par l'adoption en 2005 de la Déclaration de Paris par les pays membres du CAD ^[3]. En particulier, ces pays

[2] Même si le sommet du Millénaire est suivi du sommet de Monterey, où les bailleurs renouvellent leur engagement de consacrer 0,7 % de leur produit national à l'aide publique au développement.

[3] Engagement renouvelé depuis à Accra en 2008 et à Busan en 2011.

s'engagent alors à promouvoir la gestion axée sur les « résultats » de l'aide, en écho aux critiques fortes émises pendant les années 1990 sur l'efficacité de l'aide (à la fois dans sa mise en œuvre et dans l'atteinte de résultats tangibles en termes de développement des pays bénéficiaires).

Toutefois, il faut noter que, parallèlement à ces nouvelles orientations, le rôle et les attentes vis-à-vis de l'aide traditionnelle évoluent fortement, en lien avec son poids décroissant dans les sources de financement extérieur auxquelles ont accès les pays en développement. Ainsi, le rôle accru des fondations et des organisations non gouvernementales (ONG), le développement de la coopération Sud-Sud et l'augmentation des flux privés (IDE, investissements de portefeuille, transferts des migrants, recettes d'exportations), sous l'effet de la mondialisation des échanges, concourent à réduire le poids financier de l'aide publique au développement (APD) des pays membres du CAD (qui représentait en 2006 moins de 20 % des flux financiers extérieurs des pays en développement, contre plus de 60 % dans les années 1970, selon Frot et Santiso, 2008^[4]). Dans ce contexte, le rôle de l'aide publique évolue donc vers une fonction dite de « catalyseur » d'autres sources de financement (privées), qu'elles soient externes ou internes (Organisation de coopération et de développement économiques [OCDE], 2012a). De façon liée, en découle la recherche de financements plus innovants des bailleurs (Severino et Ray, 2010) et la volonté affichée de soutenir des interventions pilotes ou innovantes.

Dès lors, on voit que plusieurs facteurs se mêlent, qui vont se traduire par une attitude favorable sinon une demande, de la part des bailleurs vis-à-vis des évaluations d'impact : déplacement de l'attention de la communauté internationale des flux financiers d'aide et des grandeurs macroéconomiques au profit des conditions de vie effectives des populations bénéficiaires des politiques d'aide au développement ; corollairement, désir des bailleurs de renforcer et de démontrer l'efficacité de leurs interventions, notamment en termes de résultats de développement ; enfin, poids financier décroissant de l'APD traditionnelle qui appelle (de manière quelque peu paradoxale par rapport aux ambitions affichées des OMD) des ambitions renouvelées et des effets attendus à une échelle probablement plus réduite que par le passé.

[4] On note toutefois que le poids de l'APD au profit de l'Afrique subsaharienne reste très important (près de 50 % des financements extérieurs reçus en 2006).

1.2. Les évaluations d'impact : le « *big push* » des chercheurs

La forte expansion des évaluations d'impact tient également à un choc d'offre marqué des chercheurs en charge de les réaliser. En 2005, le *Center for Global Development* constitue, avec l'appui des fondations Bill et Melinda Gates et William et Flora Hewlett, un groupe de travail, composé de plusieurs économistes de renom^[5], dans le cadre de l'*Evaluation Gap Initiative*. Celui-ci produit en 2006 un rapport fondateur intitulé « *When will we ever learn ? Improving lives through impact evaluation* ». Ce rapport ne part du postulat qu'après avoir dépensé des centaines de milliards de dollars dans le domaine du développement, notamment dans les programmes sociaux, les bailleurs, les ONG et les gouvernements des pays en développement auraient une connaissance très limitée des impacts de la plupart de leurs programmes (en écho aux analyses très critiques de W. Easterley sur les effets de l'aide^[6]). Ce rapport pointe également les limites des évaluations menées jusqu'alors par les agences de développement, essentiellement centrées sur le suivi des projets et l'évaluation de leur mise en œuvre. Les limites des évaluations de projets menées par les bailleurs avaient au demeurant déjà été pointées par Mosley (1987), qui évoquait un paradoxe macro-micro : alors que les études macroéconomiques peinent à établir un impact positif de l'aide sur le développement, les études micro-économiques concluent largement à un impact positif (par exemple, 80 % des projets de la Banque mondiale sont considérés comme satisfaisants, de même que ceux de l'AFD, cf. encadré 2). Au total, ce groupe de travail propose de combler cette « *Evaluation Gap* », notamment en préconisant la création d'une structure collective en charge de favoriser le développement des évaluations d'impact (cf. section 1-3).

Encadré 2 Que peut-on apprendre des évaluations « standards » de projet des bailleurs ? L'exemple de l'AFD

L'AFD conduit depuis 2006 des évaluations quasi systématiques de ses projets dans le cadre d'un dispositif dit des évaluations décentralisées (dans le sens où le pilotage de ces évaluations est assuré par chacune des agences de l'AFD concernées).

...

[5] Dont François Bourguignon et Esther Duflo.

[6] « *The white man's burden* » (2006).

...

Ces travaux s'appuient sur différents principes, conformes aux critères de bonnes pratiques définis par le CAD, dont celui relatif à l'indépendance des évaluations, en principe garanti par le recours à des bureaux d'études externes à l'AFD. Deux études de capitalisation, destinées à tirer des leçons de ces évaluations, ont été réalisées en 2011 et 2013 : la première porte sur un portefeuille de 50 projets dont les évaluations ont été conduites entre 2007 et 2009 (Cafferini, 2011) ; la seconde porte sur un portefeuille d'environ 55 projets dont les évaluations ont été conduites entre 2009 et 2011 (Brillion, 2013).

Si la qualité globale de ces évaluations est jugée globalement satisfaisante (environ 80 % des rapports dans les deux bilans), plusieurs éléments peuvent conduire à relativiser ce constat général :

- les analyses qui y sont faites sont jugées trop descriptives et axées sur la mise en œuvre des projets, tandis que leurs effets directs ou indirects sont imparfaitement appréciés. Ainsi, pour la division de l'Évaluation de l'AFD, les évaluations décentralisées offrent, d'une manière générale, une bonne analyse de la pertinence et de la qualité de mise en œuvre des projets, mais elles apportent moins d'éléments d'appréciation sur les résultats et les impacts obtenus ;
- ces observations tiennent sans doute tant à un manque d'évaluabilité des projets qu'à une faiblesse des méthodes utilisées dans ces travaux d'évaluation. Ainsi, près de la moitié des évaluations évoquent des problèmes de conception, au niveau à la fois de la clarté des objectifs des projets (en termes de résultats attendus), de leur définition et de leur délimitation. Les évaluateurs relèvent également souvent la faiblesse des « cadres logiques » énoncés dans les projets. Les indicateurs associés à certaines variables (en termes d'objectifs/de résultats) ne sont par ailleurs pas toujours adaptés (ils sont même parfois indisponibles), et leur définition manque souvent de clarté, ce qui les rend difficilement mesurables. De plus, ils sont très rarement quantifiés, tant au niveau de leur situation de départ qu'en tant que valeur cible. Environ un quart des rapports pointent de façon liée la faiblesse des dispositifs de suivi des objectifs et des réalisations des projets. Les mêmes types de constats sont faits pour les variables d'impact, qui de fait sont rarement correctement appréhendées dans les rapports d'évaluation. À ce titre, il faut noter que les évaluations décentralisées, qui mobilisent des ressources relativement limitées^[7], s'appuient sur des méthodes simples et légères.

Enfin, on peut noter que, de façon globale, ces évaluations portent une appréciation très satisfaisante sur les projets de l'AFD.

...

[7] Une étude mobilise un budget moyen de l'ordre de 25 000 €

...

Tableau 1 *Appréciation de la qualité des projets de l'AFD sur la base des évaluations décentralisées (méthodologie CAD)*

	Pertinence	Efficacité	Efficience	Impact	Viabilité	Valeur ajoutée de l'AFD
Nombre d'évaluations	55	52	53	52	53	50
% d'appréciations satisfaisantes	93 %	83 %	70 %	87 %	58 %	86 %
% d'appréciations non satisfaisantes	7 %	17 %	30 %	13 %	42 %	14 %

NB : certains rapports d'évaluation ne comportent pas d'appréciation sur un critère.

Source : Brillon, 2013.

De façon corollaire, une « innovation » méthodologique donne une impulsion très forte aux évaluations d'impact. Il s'agit de la méthode par assignation aléatoire contrôlée (RCT), qui permet en théorie de mesurer de façon à la fois très robuste et « simple » l'effet strictement attribuable aux actions de développement. Cette innovation^[8] est à l'origine de la création d'un laboratoire de recherche, le *Abdul Latif Jameel Poverty Action Lab* (J-Pal), par Banerjee, Duflo, et Mullainathan, en 2003 (ce laboratoire est en fait apparu quelques années après une organisation proche dans son objet appelée le *Innovations for Poverty Action* [IPA]). Ce laboratoire, dont les études reposent uniquement sur la technique RCT, ambitionne non seulement de réaliser des évaluations d'impact de haut niveau académique, mais aussi de promouvoir la diffusion et l'utilisation politique de ces travaux (« *encourage real policy changes based on results* »). Pour parvenir à cet objectif, le J-Pal s'appuie à la fois sur un accent fort mis sur la communication des résultats obtenus (*Policy Briefs*, séminaires de restitution, etc.), bien au-delà des cercles scientifiques traditionnels, et sur

[8] Il faut toutefois noter que l'utilisation de la méthode RCT dans le domaine des sciences sociales est ancienne (Société française d'évaluation [SFE], 2010) puisqu'elle date du XIX^e siècle : elle a en effet été appliquée dans le domaine de l'éducation et de la psychologie dans les années 1890 et 1900 en Grande-Bretagne. Cette méthode a connu un âge d'or dans les années 1960 aux États-Unis, où elle a été appliquée dans de nombreux domaines : efficacité des programmes de propagande du ministère de la Défense, programmes pilotes dans les domaines de l'aide sociale, des mesures destinées à prévenir l'échec scolaire, des politiques contre le chômage, de la lutte contre la drogue, de la prévention de la criminalité, etc.

la réalisation de formations régulières à la méthodologie RCT, à destination des chercheurs comme des bailleurs et des gouvernements des pays partenaires. Il faut noter qu'au-delà de la question de l'évaluation et du choix de la méthode privilégiée, pour ces économistes, c'est plus fondamentalement l'échelle et l'ambition des projets de développement qui doivent être revues (« *piecemeal solutions to poor people's problems* »). Il s'agit ici d'une philosophie des « petits pas », qui s'oppose aux approches institutionnelles et d'ordre macroéconomique promues par le FMI et la Banque mondiale dans les années 1980 et 1990. En outre, cette approche, d'essence empirique, se veut fondée sur le pragmatisme, face au « sommeil dogmatique des idéologies », post-consensus de Washington (Durand et Nordmann, 2011). Elle propose ainsi la recherche de solutions ancrées dans le réel, dans les situations vécues par les individus, à une période où l'économie du développement peine à trouver un nouveau modèle théorique (cf. encadré 3). Leland (2011) analyse d'ailleurs la création du fonds *Development Innovation Ventures* (DIV) par l'*United States Aid for International Development* (USAID) en 2010 (finançant l'expérimentation d'approches innovantes, cf. section 1-3) comme le signe d'une nouvelle direction prise par l'agence américaine : celle d'un modèle de développement international décentralisé, fondé sur l'initiative individuelle (*market-based*), plus proche du modèle d'entreprise sociale apparu au cours des dernières années que du modèle historique « d'aide missionnaire ». Cette approche s'inscrit dans le concept du « *bottom of the pyramid* » (BOP), développé par Prahalad en 2004.

Encadré 3 Les méthodes promues d'évaluation des politiques publiques : un choix politique ?

Comme le notent Morra Imas et Rist (2009), de même que la notion de développement a évolué au cours du temps, celle d'évaluation du développement a connu des changements. Ainsi, loin d'être uniquement le résultat d'une décision neutre et objective, le choix des méthodes promues d'évaluation des politiques publiques (les « *gold standards* ») révèle sans doute les préférences politiques du moment :

- analyse coûts – bénéfices dans les années 1960, en lien avec la « croyance » dans la priorité à donner aux infrastructures, comme pilier du développement des pays du Sud ;
- analyse plus qualitative dans les années 1970, en lien avec la réorientation de l'aide vers les secteurs sociaux ;
- modèle macroéconomique dans les années 1980, en lien avec les programmes d'ajustement structurel (« *getting the prices right* ») ;

...

...

- analyse macroéconométrique et analyse qualitative axée sur la mise en œuvre dans les années 1990, en lien avec l'agenda sur l'efficacité de l'aide ;
- analyse d'impact dans les années 2000, en lien avec l'échelle revue (des effets) de l'aide et l'échec des modèles de pensée économique.

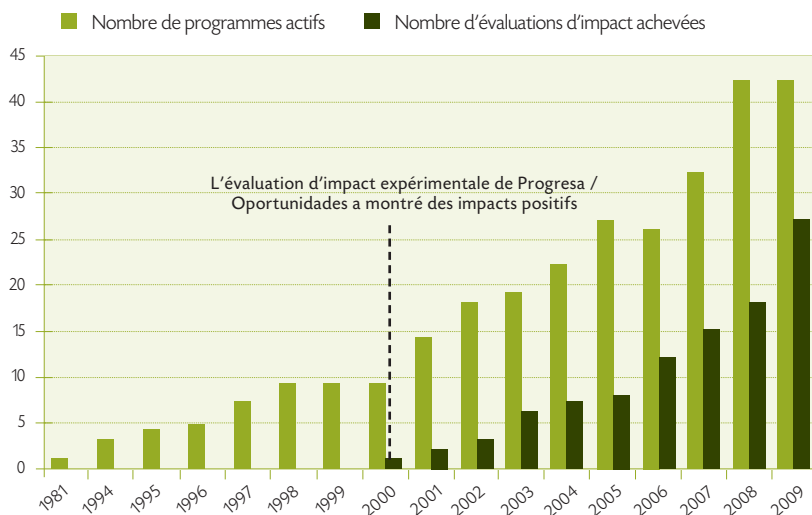
Tableau 2 Paradigme de l'aide et « gold standards » de l'évaluation

Période	Paradigme dominant de l'aide	Méthode d'évaluation promue	Discipline dominante
Années 1960	Priorité donnée aux infrastructures et à l'agriculture	Analyse coûts-bénéfices (<i>ex ante</i>)	Finance
Années 1970	Réorientation vers les secteurs sociaux (éducation, santé)	Analyse qualitative	Sciences sociales
Années 1980	Restauration du cadre macroéconomique et ajustement structurel	Modèle macroéconomique	Macroéconomie
Années 1990	« Fatigue de l'aide » : accent mis sur l'efficacité de l'aide et des critères d'allocation fondés sur la bonne gouvernance	Analyse qualitative et régressions économétriques	Macro-économétrie et sciences sociales
Années 2000	Priorité au bas de l'échelle (« <i>bottom of the pyramid</i> »), mais des ambitions contradictoires : une aide modeste (« réduire la pauvreté à petits pas ») ou une aide ambitieuse (OMD) ?	Évaluations d'impact expérimentales/ quasi-expérimentales ; gestion axée sur les résultats	Microéconomie ; statistique
Post 2015 ?	L'aide comme catalyseur /outil de politique publique globale ?	Analyses qualitative et quantitative	Pluridisciplinarité

Source : inspiré de Morra Imas et Rist (2009).

Enfin, cette forme de « révolution silencieuse » proposée par ces chercheurs (« *Just as randomized evaluations revolutionized medicine in the 20th century, they have the potential to revolutionize social policy during the 21st* » [Duflo et Kremer, 2003]) s'appuie sur l'exemplarité de l'évaluation d'impact d'un programme de prestations sociales conditionnelles au Mexique, le programme Progresa, conduite à la fin des années 1990. Ainsi, ce programme a su conjuguer tous les ingrédients d'une évaluation d'impact réussie : il s'est appuyé sur une méthodologie robuste, a fourni des résultats très positifs, et surtout s'est traduit politiquement par le maintien et l'expansion du programme au Mexique (en dépit des alternances politiques), ainsi que sa répllication dans plus de 40 pays (cf. graphique 2). Ces travaux ont fait l'objet de publications très nombreuses et sont abondamment cités comme exemple en faveur de la promotion des évaluations d'impact (cf. Bernard *et al.*, 2012)^[9]. Cette évaluation d'impact, au même titre que d'autres sur l'administration de vermifuges ou la distribution de moustiquaires imprégnées, sont d'ailleurs considérées comme des travaux influents par la Banque mondiale (2009).

Graphique 2 Prestations sociales conditionnelles : nombre de programmes et d'évaluations d'impact de ces programmes au niveau mondial



Note : Le nombre de programmes actifs est le nombre de programmes actifs cumulés (hors programmes clos). Le nombre d'évaluations d'impact est aussi un nombre cumulé.

Source : IEG World Bank (2012).

[9] Même si, dans les années récentes, ces travaux ont suscité des débats controversés sur l'interprétation à donner aux résultats obtenus et à leur utilisation (cf. section 2-3-1).

Au total, il est intéressant de noter que demande de (certains) bailleurs et offre des chercheurs (appuyés par les fondations et ONG) se rencontrent au début des années 2000 pour parvenir à un équilibre propice à une forte production d'évaluations d'impact, mais que cette rencontre semble se baser dans une certaine mesure sur un malentendu. Souci de démontrer et d'améliorer l'efficacité de leurs actions en termes de résultats et d'impacts de développement d'un côté ; tests de solutions locales aux enjeux de développement de l'autre. Néanmoins, ces deux tendances se rejoignent sur un élément clé, qui est l'attention accrue portée aux effets des interventions sur les bénéficiaires finaux.

1.3. La mise en place de structures et de financements inédits en faveur des évaluations d'impact

En réponse à l'appel des chercheurs, la « communauté des bailleurs » met en place au cours des années 2000 des moyens considérables en faveur des évaluations d'impact, via en particulier des structures dédiées, ainsi que des réseaux de praticiens. On peut citer en particulier :

- le 3IE, qui a été créé en 2008. Cette structure joue un double rôle : elle est à la fois une agence de financement et un « knowledge broker ». Ses activités incluent notamment le financement et l'appui à la conduite d'évaluations d'impact et de revues systématiques de la littérature, la constitution de bases d'évaluations d'impact et de revues systématiques de la littérature, ainsi que l'organisation de séminaires de formation et de dissémination des résultats des travaux financés/réalisés. Le 3IE joue également un rôle de définition de standards de qualité (au travers notamment de notes méthodologiques et de services d'appui technique) et d'identification des sujets prioritaires à traiter (au travers du choix des études financées et de l'établissement de « gap maps »). Il a bénéficié depuis 2008 de financements considérables (plus de 68 M USD), principalement en provenance de deux sources (le *Department for International Development* [DFID] et la fondation Bill et Melinda Gates), suivies par d'autres institutions (*Australian Agency for International Development* [AusAid], William et Flore Hewlett, etc.).
- Parallèlement, la Banque mondiale héberge successivement deux fonds dédiés au financement des évaluations d'impact : le *Spanish Impact Evaluation Fund* créé en 2007 et ayant bénéficié d'un financement de près de 14 M USD de la Coopération espagnole, et de 1,5 M USD du DFID. Il a permis de financer 51 évaluations d'impact sur la période 2007-2012, ainsi qu'un nombre très

important de formations régionales dans les pays en développement sur les évaluations d'impact. Ce fonds s'est éteint en juin 2012 et a été relayé par le *Strategic Impact Evaluations Fund* (SIEF), fonds bénéficiant d'une dotation par le DFID de près de 40 M USD. Ce fonds vise à financer environ 50 évaluations d'impact sur la période 2012 -2017 (pour un montant total de 25 M USD).

- L'USAID crée de son côté fin 2010 un fonds dédié au financement de l'expérimentation d'approches innovantes dans les pays en développement : le *Development Innovation Ventures* (DIV). Le budget annuel de ce fonds peut atteindre jusqu'à 24 M USD (limite étendue à plus de 28 M USD pour l'année 2014). Il financerait actuellement environ 60 initiatives dans huit secteurs et 22 pays. En outre, en juin 2013, l'USAID et le DFID ont annoncé la création d'un nouveau fonds (*Global Development Innovation Ventures* [GDIV]), se substituant au DIV, dans le cadre de la *National Impact Initiative*.
- Certains gros bailleurs multilatéraux se sont également investis de façon très marquée dans la conduite « en interne » de cette activité (surtout la Banque mondiale, la Banque inter-américaine de développement ; cf. section 2-1). Cette orientation s'est en particulier traduite pour la Banque mondiale par la création d'un programme transversal dédié à cette activité en 2005 (*Development Impact Evaluation Initiative* [DIME]).
- Enfin, le développement des évaluations d'impact s'est accompagné de l'apparition de réseaux de chercheurs et/ou de praticiens dédiés tels que le NONIE (*Network of Networks of Impact Evaluations*). Ce réseau, créé en 2006, regroupe à la fois des réseaux d'évaluateurs de bailleurs bilatéraux (OCDE/CAD – *Network on Development Evaluation*) et multilatéraux (*Evaluation Cooperation Group* ; *United Nations Evaluation Group*), ainsi que de pays en développement (*International Organization for Cooperation in Evaluation*). Il vise à encourager le développement d'évaluations d'impact de qualité, via la définition de standards de qualité (cf. Guide du NONIE de 2009) et les échanges de bonnes pratiques entre bailleurs, notamment au travers de conférences organisées, en principe, annuellement ^[10].

[10] À noter toutefois que l'existence de ce réseau, devenu progressivement moins actif, est actuellement remis en question.

Le soutien, notamment financier, dont bénéficient ces structures est toutefois relativement hétérogène du côté des bailleurs (voir également section 2). *A contrario*, ce soutien est très marqué du côté des fondations d'entreprises et des ONG^[11]. Ainsi, ce sont les fondations Bill et Melinda Gates et William et Flora Hewlett qui lancent l'*Evaluation Gap Initiative* en 2005 à l'origine de la forte expansion des évaluations d'impact dans le domaine du développement. Ce sont également elles qui soutiennent fortement le 3IE et le J-Pal. L'ONG *International Child Support* (ISC) « héberge » les premières évaluations d'impact expérimentales du J-Pal au Kenya, dont celle emblématique sur les vermifuges. En France, certaines fondations d'entreprises sont également très actives dans le domaine, comme par exemple la fondation Véolia ou la fondation Danone. Les frontières entre évaluateurs (notamment du J-Pal) et ONG et fondations sont d'ailleurs parfois ténues : E. Duflo appartient au Conseil d'orientation et au réseau d'experts de certaines fondations (Véolia, Danone) ; l'expérience de l'ISC en est également une bonne illustration (Roetman, 2011)^[12]. Enfin, il est même arrivé que l'IPA se retrouve en position d'opérateur de programmes de développement, en dehors de tout cadre d'évaluation (par exemple, pour étendre géographiquement certains dispositifs auparavant évalués).

1.4. Les évaluations (d'impact)... au-delà de l'aide

Il faut enfin noter que le développement des évaluations d'impact au cours de la dernière décennie est loin d'être spécifique au seul secteur du développement. Dans le contexte de crise financière et économique post 2008, la plupart des gouvernements des pays développés sont engagés dans des politiques de réduction des dépenses publiques, où les évaluations (d'impact) jouent un rôle croissant. En effet, il s'agit de cibler les coupes budgétaires et de réorienter les ressources publiques vers les politiques les plus efficaces. En France, malgré un certain retard par rapport à d'autres pays, on note, depuis plusieurs années, que l'évaluation des politiques publiques s'est imposée progressivement comme une priorité forte du gouvernement, comme outil d'accompagnement et d'orientation de ses réformes. On peut citer l'étude expérimentale emblématique sur le Revenu de solidarité active (RSA)^[13] (conduite en

[11] Pour Rosenzweig (2012), il n'est pas surprenant que ces travaux séduisent (notamment) les chercheurs et les ONG : « *The message that economic science and money (aid) can be combined to directly help the poor, even in corrupt and/or dysfunctional political regimes and in the absence of grand strategies or major technological breakthroughs, is of obvious appeal to altruistic billionaires, aid agencies, charitable NGOs, and dewy-eyed college economics majors who see that they can both do good and learn in exotic locales* ».

[12] « *The line between programmes and evaluations began to blur* » ; « *The choice of research questions had implications for programmatic choices and vice versa* ».

[13] Il faut noter que si cette évaluation est expérimentale (au sens où l'on teste différentes modalités de mise en œuvre d'une politique), il ne s'agit toutefois pas d'une expérimentation *aléatoire* contrôlée (les zones/publics bénéficiaires n'étant pas sélectionnés par assignation aléatoire).

2007-2008) et la création en 2008 d'un fonds d'expérimentation pour la jeunesse, destiné à tester l'efficacité d'un certain nombre de politiques (visant à favoriser la réussite scolaire des élèves et l'insertion sociale et professionnelle des jeunes de moins de vingt-cinq ans) avant passage à l'échelle^[14]. Récemment, le CEPII (Crozet *et al.*, 2013) a également réalisé une évaluation d'impact (quasi-expérimentale) des dispositifs de soutien aux exportations françaises pour le compte du ministère de l'Économie.

En outre, tous les projets de loi (sauf exceptions prévues par une loi organique) doivent être accompagnés depuis 2009 d'une étude d'impact (*ex ante*), réalisée par le gouvernement, définissant les objectifs poursuivis, exposant les motifs du recours à une nouvelle législation, l'état actuel du droit dans le domaine visé, l'articulation du projet avec le droit européen, l'évaluation des conséquences économiques, financières, sociales et environnementales des dispositions du projet et les modalités d'application envisagées ainsi que leurs conséquences.

Dernière pierre à cet édifice, la modernisation de l'action publique (MAP) annoncée en décembre 2012 prévoit désormais que « toutes les politiques publiques, sur l'ensemble du quinquennat, feront l'objet d'une évaluation » (voir discours de J.-M. Ayrault du 18 décembre 2012). Ainsi, une première vague d'une quarantaine d'évaluations est prévue sur des politiques dans des domaines variés (tels que les aides aux entreprises, les aides au logement, l'éducation, etc.), représentant plus de 20 % de la dépense publique annuelle. Cette volonté s'appuie notamment sur les travaux du Conseil d'analyse économique (CAE, 2013).

De façon générale, la place et la culture de l'évaluation des politiques publiques *domestiques* des pays donateurs influencent très certainement la place plus ou moins forte accordée aux évaluations d'impact, et plus généralement à la gestion axée sur les résultats, dans leur politique internationale de développement.

[14] Parmi l'ensemble des expérimentations lancées par ce fonds depuis juillet 2009, un peu plus de 55 % impliquent un protocole d'évaluation basé sur une comparaison entre un groupe témoin et un groupe cible, dont une partie (représentant environ 37 % du total) avec assignation aléatoire (Gurgand et Valdenaire, 2012).

2. *What have we learnt?*

Points de consensus atteints au niveau international^[15]

2.1. « *Size matters* » ?

Bien que 57 % des bailleurs déclarent réaliser des évaluations d'impact (OCDE, 2010), un examen plus détaillé met en évidence une grande hétérogénéité parmi les bailleurs. Deux bailleurs multilatéraux (la Banque mondiale et la Banque inter-américaine de développement) se distinguent par une production à la fois très importante et relativement « ancienne » d'évaluations d'impact de leurs programmes. Pour exemple, la Banque mondiale lance depuis 2005 plus de 45 évaluations d'impact par an^[16]. Un troisième bailleur, cette fois-ci bilatéral (le DFID), est également relativement atypique en ayant contribué de façon très significative au *financement* des évaluations d'impact, *via* des fonds ou des structures collectives (le SIEF, le 3IE, par exemple), tout en n'ayant jusqu'à récemment pas conduit directement de tels travaux (voir DFID, 2013)^{[17],[18]}.

[15] Les éléments présentés dans cette partie se basent à la fois sur les informations collectées sur les sites des principaux bailleurs identifiés (dont de nombreuses publications), et, lorsque cela a été possible, sur des échanges directs avec les services d'évaluation de ces institutions (voir annexes 1 et 2). L'analyse porte sur les principaux bailleurs internationaux en termes de taille.

[16] Chiffre établi à partir de la base en ligne de la Banque mondiale. Selon l'*Independent Evaluation Group* (IEG) (2012), ce chiffre atteindrait même 62.

[17] Cela a été le cas également de la coopération espagnole qui a toutefois arrêté, depuis 2012, ses financements à de tels fonds.

[18] L'USAID est dans une certaine mesure dans une situation similaire.

Parmi les principaux autres bailleurs, on distingue ensuite trois groupes :

- les bailleurs dont la production d'évaluations d'impact est significative (au moins deux à trois évaluations d'impact lancées par an) comme le *Millenium Challenge Corporation* (MCC), l'*Asian Development Bank* (ADB), la *Japan International Cooperation Agency* (JICA), l'*USAID* et la coopération néerlandaise. Ces bailleurs se caractérisent par un accent relativement récent mis sur ce type d'évaluation, mais aussi par une orientation stratégique affirmée en la matière^[19] ;
- les bailleurs qui réalisent des évaluations d'impact de façon « opportuniste » ou encore exploratoire comme l'AFD, la KfW et le GIZ. Ces bailleurs se caractérisent par un nombre beaucoup plus restreint de travaux réalisés (pas plus d'une étude par an lancée depuis 2005), et une activité qui reste considérée comme pilote (i.e. non institutionnalisée) ;
- les bailleurs qui ne réalisent pas d'évaluations d'impact, et ne devraient pas à court terme fortement investir dans cette activité. C'est le cas en particulier des institutions de l'Union européenne (Commission européenne, Banque européenne d'investissement (BEI)), qui évoquent principalement des barrières institutionnelles et financières (coût souvent élevé de ces études) au développement de cette activité.

On peut enfin distinguer le cas des agences des Nations unies, qui ont réalisé jusqu'à présent peu d'évaluations d'impact (au sens strict du terme), mais ont constitué une *task force* spécifique sur cette question en 2009 (*United Nations Evaluation Group* [UNEG], *Impact Evaluation Task Force* [IETF]), qui a abouti en 2013 à la réalisation d'un guide sur les évaluations d'impact, spécifique aux activités des Nations unies, et visant à favoriser la réalisation de telles études, sous contrainte néanmoins d'un budget disponible qualifié de limité.

[19] Le MCC accorde une place privilégiée aux évaluations d'impact (avec un budget compris entre 1 et 2 M USD par étude dans le domaine de l'agriculture par exemple). La Banque asiatique de développement a impulsé cette activité sous la forme d'une initiative lancée en 2010 (*Implementing Impact Evaluations at ADB*), appuyée par une assistance technique. La JICA, après une phase pilote, a également entrepris d'institutionnaliser ce type d'évaluations, grâce notamment à l'institut de recherche créé en son sein en 2008 (*JICA Research Institute*) et à des financements conséquents (le budget moyen d'une étude peut varier entre 100 000 USD et 1 M USD). L'USAID a récemment inscrit les évaluations d'impact comme une forme d'évaluation relativement importante dans son document de politique d'évaluation (2011). Elle a fixé un minimum d'une évaluation d'impact par « objectif de développement » à identifier (au sein de ses documents fixant les stratégies pays), et collabore également avec le 3IE pour lancer des revues systématiques de la littérature. La coopération néerlandaise met depuis plusieurs années un accent croissant sur les « méthodes statistiques d'évaluations d'impact », et y consacre des ressources conséquentes (par exemple, elle a lancé récemment un appel d'offre pour une série de cinq évaluations d'impact dans le domaine des énergies renouvelables pour un budget global de 5 M€), parallèlement au mouvement de concentration géographique de son aide (de 33 pays d'intervention à 15 d'ici 2015).

Tableau 3 Une corrélation imparfaite entre taille des bailleurs, concentration des engagements et adhésion aux évaluations d'impact

Bailleurs	Engagements bruts (Md USD, 2011)	Nombre de pays d'intervention	Ratio engagements/ nombre de pays (M USD, 2011)	Nombre d'évaluations d'impact (achevées, en cours et prévues)*
Groupe 1 : les gros producteurs				
Banque mondiale :				
• IBRD	26,7	188	142	483 (au moins)
• IDA	16,3	81	201	
IDB	11,4	48	238	100 (au moins)
Groupe 2 : les gros financeurs				
DFID	6,8	78	87	Financement conséquent fonds
Groupe 3 : les producteurs moyens				
MCC	1,4	25	56	48 (dont 40 prévues)
AsDB	21,7	67	324	37
JICA	19,9	95	209	26
USAID	11,2	140	80	44 (prévues)
Coopération néerlandaise	4,2	33	128	≈ 20
Groupe 4 : les petits producteurs				
AFD	8,0	73	110	10
KfW	6,3	140	65	≈ 10
GIZ	2,8	130	54	≈ 5
Nations unies	11,3	147	77	≈ 10
Groupe 5 : les non producteurs				
UE (CE, BEI)	12,3	150	82	-
AfDB	8,5	53	160	-

* Les chiffres mentionnés ont été établis à partir des informations recueillies auprès des différents bailleurs (soit via leur site/leurs publications, soit directement auprès des services d'évaluation). Ils doivent donc être considérés avec précaution.

Sources : rapports, sites et entretiens auprès des bailleurs mentionnés.

Force est de constater qu'une part considérable des évaluations d'impact n'est en fait pas *directement* commanditée par les bailleurs de fonds sur leurs propres programmes :

- une grande partie est conduite, après appel à projets, grâce à l'appui de fonds multilatéraux financés à la fois par certains bailleurs et fondations : la somme cumulée des évaluations d'impact financées par le 3IE, le SIEF (I et II) et le DIV atteindrait ainsi près de 300 ;
- corollairement, une partie importante de ces travaux est le fruit d'un partenariat direct entre chercheurs et maîtrises d'ouvrage/opérateurs (ONG et, dans certains pays, ministères), avec financement ou non des bailleurs. Le J-Pal/IPA et le *Center for Effective Global Action* (CEGA), qui constituent des centres de recherche leaders en matière d'évaluations d'impact, totaliseraient près de 500 évaluations d'impact (dont une partie réalisée directement avec les maîtrises d'ouvrage nationales ^[20]).

Tableau 4 Les évaluations d'impact vues du côté des fonds multilatéraux et des centres de recherche leaders

Institutions	Budget global (M USD)	Nombre d'évaluations d'impact (achevées, en cours et prévues)
Selon les fonds multilatéraux		
3IE	68	129
SIEF I	15,5	51
SIEF II	40 (25 dédiées au financement des EI)	50
DIV	≈ 20	≈ 60
Selon les centres de recherche leaders		
J-Pal	nd	397
CEGA	nd	Environ 100

Sources : rapports et sites des institutions mentionnées.

[20] Les bases recensant les évaluations d'impact de ces institutions ne permettent toutefois pas aisément d'extraire cette information. En outre, ce chiffre global inclut en partie des doublons compte tenu des financements croisés et des affiliations parfois multiples des chercheurs.

Au total, cette première typologie de la production d'évaluations d'impact permet de faire ressortir deux éléments clés. D'une part, l'essentiel des évaluations d'impact réalisées au niveau international émane, d'un côté, de certains gros bailleurs (Banque mondiale, *Inter-American Development bank* [IDB]), de l'autre, de fonds multilatéraux et de collaborations directes entre chercheurs et maîtrises d'ouvrage locales. Dans ce cadre, le modèle traditionnel des évaluations (d'impact) *ex post* de projet semble s'effacer au profit d'un modèle, où les fonds multilatéraux jouent un rôle important (cf. schémas 1 et 2). Cette « prime à la taille » n'est pas surprenante compte tenu des contraintes inhérentes à la réalisation des évaluations d'impact (budget conséquent, temps de réalisation souvent long, importance des ressources humaines à mobiliser ; cf. encadré 4). De plus, les interactions fortes entre *design* du projet et *design* de l'évaluation s'agissant de la méthode expérimentale (cf. annexe 3) expliquent la nécessité de liens beaucoup plus étroits entre chercheurs et opérateurs des projets que dans le cas des évaluations (d'impact) *ex post* traditionnelles (« [...] *it makes the researcher transition from spectator to actor in the development process* », Legovini, 2010).

Schémas 1 et 2 La gouvernance des évaluations d'impact : la fin du modèle principal-agents

Schéma 1. Modèle traditionnel des évaluations (d'impact) *ex post* chez les bailleurs

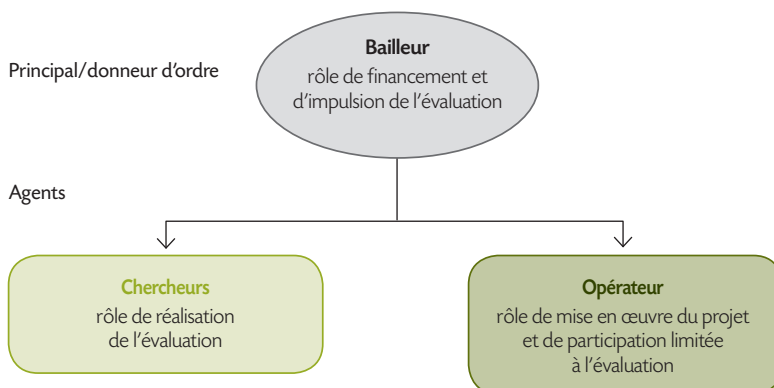
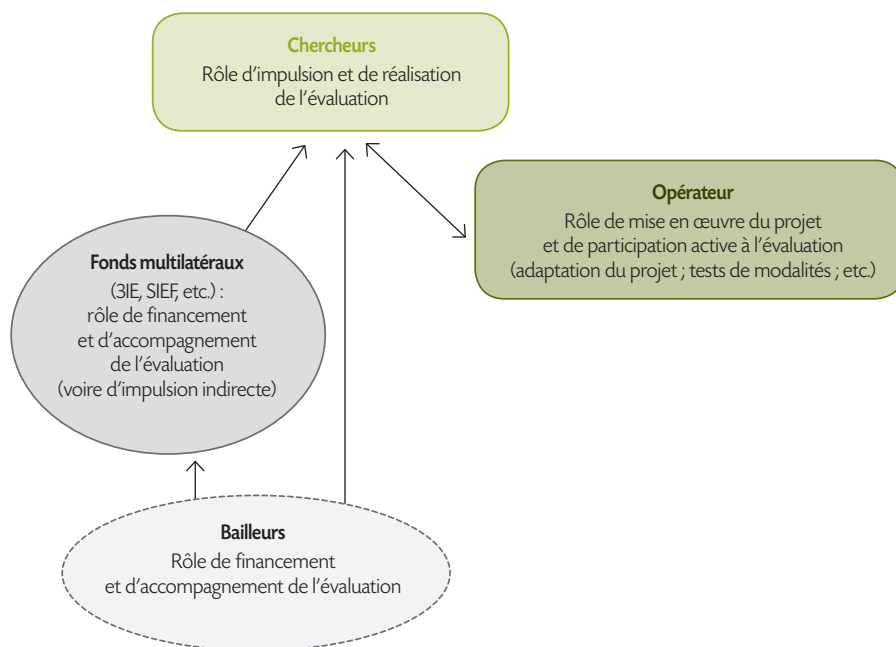


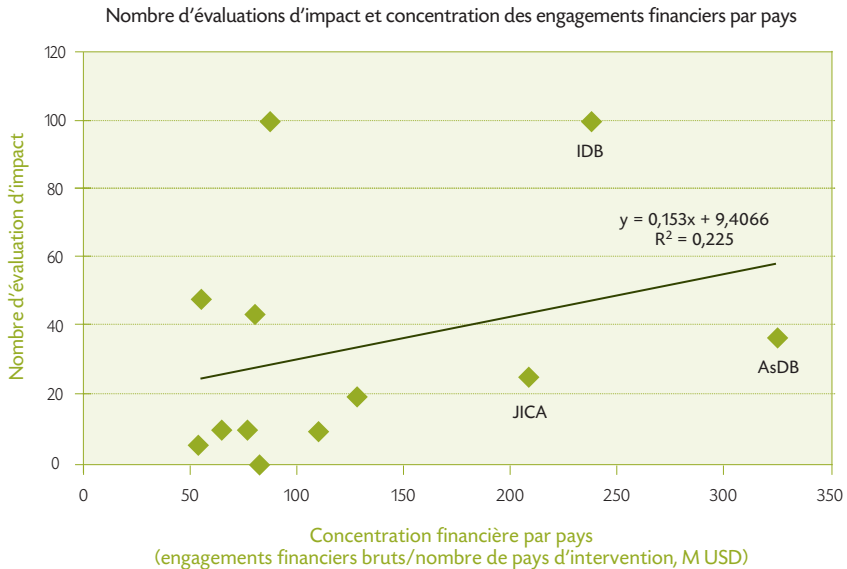
Schéma 2. Un exemple de modèle d'évaluations d'impact prospectives



D'autre part, parmi les bailleurs qui conduisent des évaluations d'impact de leurs propres programmes (hors Banque mondiale), seul un lien ténu entre taille (au sens du montant des engagements financiers totaux des bailleurs ou rapportés au nombre de leurs pays d'intervention) et production d'évaluations d'impact peut être établi, ce qui signifie que d'autres facteurs sont plus déterminants dans l'appétence des bailleurs pour internaliser cette activité (culture de l'évaluation ; probablement type d'instruments financiers dominants^[21] ; etc.).

[21] Il n'est toutefois pas aisé d'établir une relation entre le type d'aide délivrée par les bailleurs et la place qu'ils accordent aux évaluations d'impact (par exemple, l'aide projet est censée mieux se prêter à ces études), compte tenu de la fragilité des données du CAD.

Graphique 3 La taille des engagements des bailleurs (hors Banque mondiale) par pays : un prédicteur faible de leur activité d'évaluations d'impact



Sources : Bailleurs, calculs auteurs.

Encadré 4 L'économie des évaluations d'impact : éléments chiffrés à partir de l'expérience de quelques bailleurs

Budget

Du côté des bailleurs (qui évaluent leurs programmes)

Le budget moyen consacré à une évaluation d'impact par la Banque mondiale est de 500 000 USD (IEG, 2012). Ce budget élevé doit toutefois être mis en rapport du montant moyen des projets financés par la Banque mondiale : il n'en représente ainsi « que » 1,4 % en moyenne. MCC et l'USAID évoquent des budgets moyens plus conséquents (de l'ordre de 1 à 2 M USD – dans le domaine de l'agriculture), qui s'expliquent par l'externalisation totale de ces études (alors que la Banque mondiale n'externalise « que » la collecte de données). Le budget moyen des évaluations d'impact conduites par la KfW (budget partagé – pour certaines – avec d'autres bailleurs) est plus modéré,

...

...

de l'ordre de 200 000 €, sans doute en lien avec un recours plus systématique à des données existantes (et à une internalisation partielle). La JICA évoque quant à elle une fourchette budgétaire large, variant de 100 000 USD à 1 M USD par étude, rejoignant par-là l'AFD ayant supporté un coût unitaire s'étalant de 35 000 € à 800 000 K€ (cofinancé) selon les évaluations.

Seule l'IDB semble se distinguer des autres bailleurs : ainsi, dans le bilan qu'elle faisait en 2008 de ses évaluations d'impact (Ruprah, 2008), elle évoquait un coût moyen par étude modéré (43 000 USD), expliqué par trois raisons principales : le recours prioritaire à des données existantes ; une stratégie d'évaluation simultanée du même programme dans plusieurs pays, afin de « rentabiliser » le protocole ; le recours important à des consultants/chercheurs locaux via le réseau *Evalnet*. Elle notait néanmoins les limites de cette stratégie, à savoir un choix des études réalisées guidé par la disponibilité des données, conduisant à un taux d'échec très important (65 % des projets d'études).

Du côté des fonds multilatéraux

Le montant moyen accordé par le SIEF I au financement d'une évaluation d'impact peut être estimé à environ 300 000 USD, proche de celui évoqué par le 3IE (AusAid, 2011) de l'ordre de 250 000 USD et similaire à celui mesuré pour le DIV (300 000 USD, variant de 75 000 USD à 1,5 M USD, selon les études). À noter que le financement accordé à une étude via ces fonds peut venir dans certains cas s'ajouter à d'autres sources de financement (dans le cas du DIV, il s'agit même d'un prérequis).

Au total, si ces chiffres sont bien évidemment parcellaires, il ressort que le budget nécessaire au financement d'une évaluation d'impact est généralement conséquent, même s'il existe une forte variabilité en fonction des données mobilisées, du mode de réalisation de ces études (interne/externe ; local/international), ainsi que de la méthode employée (dans le cas de la méthode expérimentale, la surveillance du respect de l'intégrité du protocole génère un surcoût – qualifié de marginal toutefois par Diamond (2012) – par rapport à d'autres approches). Il faut toutefois relativiser ces chiffres par rapport au budget moyen consacré à d'autres types d'évaluation : dans le cas de l'AFD, le coût moyen des évaluations dites stratégiques et thématiques (qui reposent généralement sur des approches plus qualitatives, mais dont le champ d'analyse est plus large que les évaluations d'impact) est estimé à environ 125 000 €. En outre, les évaluations d'aide budgétaire menées ont un coût moyen de l'ordre de 400 000 € (cofinancées avec d'autres bailleurs).

Temps de réalisation et incertitudes

Le temps de réalisation d'une évaluation d'impact est généralement long. Dans le cas de l'AFD, le temps de préparation d'une étude de type prospectif (c'est-à-dire dont la conception intervient avant la mise en œuvre du projet et pour laquelle une collecte

...

...

de données de référence est nécessaire) est estimé à 1 an/1 an et demi. À cela s'ajoute aisément un à trois ans de déploiement (en fonction de la date à laquelle est réalisée l'enquête *ex post*), puis un à deux ans minimum d'exploitation des données et de diffusion des résultats. Ainsi, dans le cas d'une étude expérimentale menée, les premiers travaux d'identification et de faisabilité de cette étude ont débuté en 2005 et la restitution finale des résultats obtenus a eu lieu en 2013 ! D'autres études ont toutefois pu être menées à leur terme dans des délais beaucoup plus courts : par exemple, une étude quasi-expérimentale, reposant sur des données existantes, a pu être conduite de bout en bout en environ 2 ans.

À cette durée souvent longue de réalisation, s'ajoute un élément d'incertitude fort sur l'aboutissement des travaux engagés et, dans le cas des études expérimentales, sur la capacité à préserver l'intégrité de la méthode sur une période de temps assez longue. Pour illustration, dans le cas de l'AFD, sur une vingtaine d'études sur lesquelles des travaux ont été engagés (au minimum une analyse de faisabilité jusqu'au maximum une enquête de référence), dix ont dû être abandonnées (pour des raisons multiples allant d'un manque d'évaluabilité du projet à un problème de financement ou à un partenaire défaillant) et trois en cours ou prévues présentent des risques importants de ne pas aboutir (soit un taux d'échec d'environ 45 % à 60 %, un peu inférieur à celui évoqué par l'IDB). De façon corollaire, les cas d'expérimentations « cassées » mis en évidence par la Banque mondiale (Campos *et al.*, 2012) et le MCC (2012) témoignent de la fragilité de ce type d'études. Dans le cas des deux expérimentations menées par l'AFD, une vague d'enquêtes *ex-post* a dû être supprimée compte tenu de difficultés opérationnelles (taux d'adhésion restant bas dans les deux cas ; taux de résiliation des contrats élevé s'agissant de la micro-assurance de santé).

2.2. Du « *gold standard* » aux « *good standards* »

2.2.1. De la « bulle » expérimentale...

L'expérimentation aléatoire contrôlée (cf. encadré 5) a été présentée, à partir du milieu des années 2000, comme « l'étalon-or » des méthodes d'évaluation d'impact dans le domaine du développement, que cela soit dans certains travaux académiques (Duflo *et al.*, 2006) ou dans les standards d'évaluation de certains bailleurs (notamment USAID ^[22], DFID ^[23]). De fait, il apparaît clairement que la Banque mondiale, certains

[22] « For impact evaluations, experimental methods generate the strongest evidence. Alternative methods should be utilized only when random assignment strategies are infeasible » (USAID, 2011).

[23] « Experimental methods are uniquely capable of establishing causality with a high degree of statistical confidence, dealing with selection bias and providing rigorous estimates of the size of impact » (DFID, 2013).

bailleurs bilatéraux et les fonds multilatéraux (3IE, SIEF, DIV) ont accordé une place privilégiée (et croissante) à cette méthode particulière d'évaluations d'impact (cf. tableau 5 et graphique 4). Or, si faire appel à des méthodes quantitatives robustes pour mesurer les effets *attribuables* aux actions des bailleurs paraît naturel (par rapport à des méthodes qualitatives), la polarisation sur la méthode RCT semble en décalage avec la réalité des interventions évaluables selon cette méthode. Ainsi, seuls environ 5 % des interventions des bailleurs (tels que le DFID) présenteraient des caractéristiques (cf. annexe 3) se prêtant à l'utilisation d'une RCT (Stern *et al*, 2012). D'ailleurs, la majorité des bailleurs n'a de fait conduit jusqu'à présent qu'une minorité d'expérimentations aléatoires contrôlées sur leurs propres projets (celles-ci représentant généralement moins d'un quart des évaluations d'impact réalisées ou en cours)^[24]. De plus, s'agissant de la Banque mondiale, l'importance des expérimentations aléatoires contrôlées s'explique par l'activité du programme DIME, qui est en fait une utilisation très opérationnelle (*implementation science* ; voir section 2-3-4). Au total, la forte production de RCT dans le domaine du développement depuis une petite dizaine d'années semble traduire davantage une logique de l'offre (forte promotion de cette méthode par les chercheurs en lien avec des ONG), qu'une logique de la demande (de la part des bailleurs traditionnels).

[24] Même si ce chiffre est en soi déjà finalement élevé, compte tenu de la faible applicabilité de cette méthode au portefeuille standard des bailleurs !

Tableau 5 Importance de l'utilisation de la méthode RCT (par institution)

Institutions (classées par ordre décroissant)	Méthode RCT (% des évaluations totales)	Commentaires
Bailleurs		
Banque mondiale	80 (études en cours) 30 (études achevées)	% récent élevé lié à l'utilisation prospective de l'évaluation par DIME
MCC	63	Sur la base des huit évaluations d'impact achevées
JICA	50 (études en cours / prévues) 0 (études achevées)	Majorant
AFD	27	Correspondant à 2 RCT
AsDB	25 (depuis 2010) 12 (jusqu'en 2009)	–
IDB	13	Rapport 2012
GI	13	Sur la base des évaluations recensées dans GIZ (2012)
KfW	10	Correspondant à 1 RCT en cours
Coopération néerlandaise	0	Sur la base des évaluations recensées sur son site
Fonds multilatéraux		
3IE	74	Sur la base des évaluations d'impact financées par le 3IE
SIEF I	74	Critères de sélection des projets favorisant cette méthode (score technique maximum)
SIEF II	nd	Assouplissement des critères de sélection en faveur des méthodes quasi-expérimentales
DIV	64	–
Centres de recherche leaders		
J-Pal	100	Seule méthode utilisée
CEGA	54	–

* Les chiffres mentionnés ont été établis à partir des informations recueillies auprès des différents bailleurs (soit via leur site, soit directement auprès des services d'évaluation). Ils doivent donc être considérés avec précaution.

Sources : institutions mentionnées.

Graphique 4 Montée en puissance des évaluations expérimentales à la Banque mondiale



Note : Basé sur 381 évaluations d'impact lancées à la Banque mondiale (sur 396) entre 2000 et 2010.

Sources : IEG, 2012.

Encadré 5 Focus sur la méthode expérimentale ou RCT

Au cœur de l'expérimentation aléatoire contrôlée, se trouve le principe du tirage aléatoire d'individus (ou d'unités statistiques) bénéficiaires d'un programme au sein d'une population cible. Basé sur la loi des grands nombres, le tirage aléatoire permet de garantir qu'ex ante bénéficiaires et non-bénéficiaires d'un programme seront en moyenne identiques (à la fois en termes de caractéristiques observables – comme l'âge, le revenu, l'éducation par exemple – et de caractéristiques non observables – comme la motivation, l'aversion au risque par exemple). Ce faisant, cette méthode permet de garantir en théorie l'élimination du biais de sélection (ou de l'effet de composition) et donc de s'assurer que la différence constatée sur des variables de résultat ou d'impact entre groupe de bénéficiaires et groupe de non-bénéficiaires sera uniquement attribuable à l'intervention.

...

...

À noter qu'il existe différentes modalités de mise en œuvre du principe de tirage aléatoire – au-delà de la « simple » loterie : le phasage – aléatoire – dans le temps des bénéficiaires (*pipeline approach*), utilisé notamment quand un programme se déploie sur la quasi-totalité d'un territoire ; le système d'incitation/de promotion aléatoire (*encouragement design*), qui permet de ne pas exclure *a priori* des populations d'un programme et s'avère utile quand le taux d'adhésion à un programme est « naturellement » plutôt faible. En outre, il existe également des cas d'expériences dites naturelles où les conditions particulières d'un programme permettent de se rapprocher du cadre d'une RCT : c'est le cas par exemple des programmes pour lesquels il existe des seuils stricts – et exogènes – d'attribution (par exemple, une allocation attribuée en-deçà d'un certain seuil de revenu) ou des frontières géographiques. Dans ce cas, la comparaison des individus au voisinage de ce seuil/de cette frontière fournit en principe des estimations très robustes de l'impact *local* du programme (approche dite de la régression par la *discontinuité*).

Le terme de « contrôlée » fait référence au fait que, pour être valide, cette méthode nécessite que la mise en œuvre de l'expérimentation soit parfaitement maîtrisée : bonne identification de la population cible ; tirage effectivement aléatoire ; capacité à suivre les individus bénéficiaires/non-bénéficiaires tout au long de l'expérimentation ; interférence de programmes ou de chocs exogènes sur les deux groupes : minimale ou parfaitement symétrique (*i.e.* les affectant de façon similaire) ; etc. On perçoit déjà bien les difficultés de ce caractère « contrôlé » dans le cas des expérimentations sociales. Corollairement, on comprend également la nécessité de disposer de données « de première main » dans ce type de travaux, c'est-à-dire le plus souvent collectées spécifiquement pour les besoins de l'étude (données dites primaires). En effet, il est rare dans ces études de type clinique que des données existantes (dites secondaires) puissent être directement mobilisées, surtout dans le cas des pays en développement (dont la couverture de l'appareil statistique est limitée) : nécessité que ces données portent spécifiquement sur les échantillons de bénéficiaires/non-bénéficiaires sélectionnés, qu'elles aient été collectées aux dates souhaitées (soit juste avant la mise en œuvre du programme et un à trois ans après), enfin, qu'elles portent précisément sur les variables d'intérêt étudiées.

Enfin, le terme « d'expérimentation » renvoie à la logique d'essai-erreur/de test en principe propre à ce type de méthodes : il s'agit en effet d'expérimenter des programmes dont on ne connaît pas *a priori* les effets (ou tout au moins la taille) et qui pourront être, le cas échéant, revus plus ou moins fortement suite à ces tests. Bien sûr, dans le cas de projets de développement et plus généralement de politiques publiques, ce lien « simple » entre test et décision est loin d'aller de soi.

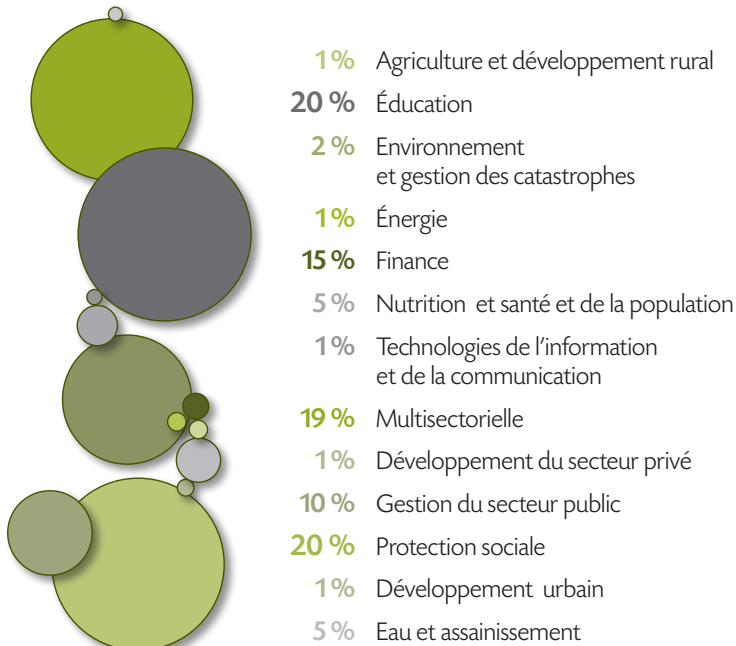
Cette « préemption » de la méthode expérimentale sur l'objet étudié semble s'être traduite par un certain ciblage des évaluations d'impact :

- les travaux réalisés semblent d'avantage porter sur de l'aide projet, que sur d'autres formes d'aide dont l'affectation à un objet précis est moins directe ; il est d'ailleurs notable de constater que l'un des bailleurs octroyant le plus de financements sous forme d'aide budgétaire (la Commission européenne) n'a pas jusqu'à présent engagé de travaux d'évaluations d'impact (expérimentales) ;
- ils concernent plutôt des interventions à échelle locale (cf. contrainte du groupe de contrôle) ;
- les évaluations d'impact expérimentales semblent également plus souvent porter sur des projets mis en œuvre par des ONG, que par des administrations/ des gouvernements nationaux (Ravallion, 2012)^[25]. Ce dernier constat s'explique par la flexibilité des opérateurs requise pour la réussite de ce type de travaux prospectifs, nécessitant des aménagements opérationnels dans la mise en œuvre « standard » des programmes. Roetman (2011) note également que l'acceptabilité du principe de l'assignation aléatoire des bénéficiaires est plus grande du côté des ONG que des gouvernements, dans la mesure où ces premières ne sont pas supposées offrir des biens et des services collectifs à l'ensemble de la population ;
- une certaine concentration sectorielle et thématique des évaluations d'impact conduites est également constatée : ainsi, dans la base du 3IE, la protection sociale (*Conditional Cash Transfers* – CCT), l'éducation et la finance (*microfinance*) représentent plus de la moitié des évaluations d'impact recensées ; dans celle du J-Pal, la finance (*microfinance*), l'éducation et la santé concernent les deux tiers des expérimentations répertoriées. La Banque mondiale note toutefois une certaine diversification des thèmes traités depuis 2005 (où les études sur le développement local, l'éducation et les CCT étaient dominantes). Il est également admis que les études expérimentales se prêtent davantage à des projets de complexité faible à modérée (NONIE, 2009 ; Stern *et al*, 2012 ; Bernard *et al*, 2012 par exemple). Ainsi, la Banque mondiale constate que les interventions complexes sont celles bénéficiant du nombre le plus limité d'évaluations d'impact, comme par exemple les projets dans les secteurs de l'énergie, du transport et des réformes institutionnelles (qui ne représentent qu'environ 6 % des évaluations d'impact de la Banque). De façon corollaire, les interventions

[25] Le manque de bases *user-friendly* (ou complètes) d'évaluations d'impact par les principaux producteurs / financeurs empêche toutefois une analyse parfaitement objective de ce dernier point.

ciblées par ces études semblent majoritairement des interventions *soft* (fourniture de services, de biens légers), plutôt que des interventions *hard* (comme celles liées aux grands projets d'infrastructures par exemple ; Ravallion, 2012)^[26], ou que celles visant à améliorer le fonctionnement des institutions des pays en développement. L'AusAid (2011) note ainsi que la majorité des évaluations d'impact financées par le 3IE, à travers son 3^e *Open Window*, concernent la fourniture de services (20 études sur 22). Enfin, une enquête réalisée par l'IEG (2012) au sein de la Banque mondiale montre que 50 % des chefs de projet et de départements estiment que les évaluations d'impact ont concerné les interventions qui étaient les plus « faciles » à évaluer, et pas forcément celles portant sur des interventions prioritaires ou pour lesquelles il y avait le plus grand manque de connaissances.

Graphique 5 Une certaine concentration sectorielle des évaluations d'impact



Source : 3IE (2011).

[26] Pour McKenzie (2011b) toutefois, il n'y aurait pas d'effet de substitution, mais bien un accroissement global de la production scientifique (quoique concentrée sur les secteurs évoqués).

Dès lors, il n'est pas surprenant qu'à une première phase d'engouement suscité par l'expérimentation aléatoire contrôlée, a succédé une période plus troublée où de vives critiques se sont fait jour, sur les limites de cette approche à la fois en termes d'utilisation opérationnelle et stratégique, comme instrument de redevabilité (cf. section 2-3), mais aussi sur le terrain méthodologique. Ainsi, un certain nombre de chercheurs remettent en cause la supériorité méthodologique de cette méthode : pour Deaton (2010) par exemple, la méthode RCT ne doit pas être considérée comme un « étalon-or » en matière d'évaluations d'impact (« *Randomization is not a gold standard because there is no gold standard* »). Ces éléments critiques, bien documentés dans la littérature, peuvent être résumés en quelques points :

- Certains auteurs pointent d'abord les difficultés pratiques de mise en œuvre des expérimentations. Ainsi, le manque de faisabilité pratique de l'assignation aléatoire simple par loterie conduit souvent à utiliser des techniques alternatives, sortes de « *second-best* », qui ne sont pas sans poser certains problèmes : phasage dans le temps qui exige une application stricte du plan de mise en œuvre graduelle prévue (MCC, 2012) ; promotion aléatoire, qui exige des comportements d'adhésion/de non-adhésion parfaitement conformes aux attentes (Diamond, 2013). Pour Labrousse (2010), la méthode RCT n'est pas exempte d'un certain « bricolage » familier des protocoles de recherche. Quentin et Guérin (2013) font le même constat sur la base de l'étude expérimentale de la micro-assurance de santé SKY au Cambodge cofinancée par l'AFD. Enfin, les cas de RCT « cassés », notamment liés à des problèmes d'adhésion insuffisante et d'autres problèmes opérationnels, semblent relativement fréquents (Campos *et al.*, 2012). Dans ces cas, la méthode effectivement utilisée pour mesurer les impacts se rapproche davantage d'une évaluation économétrique de type quasi-expérimental, que d'une RCT. Corollairement, il n'est alors pas exclu qu'en pratique, en dépit du tirage aléatoire, un certain biais de sélection persiste : cas d'attrition non aléatoire de l'échantillon^[27] ; risque de probabilité d'encouragement *de facto* non parfaitement aléatoire ; externalités générées par le programme non anticipées ; chocs/programmes exogènes affectant un seul groupe sans que cela n'ait été prévu.

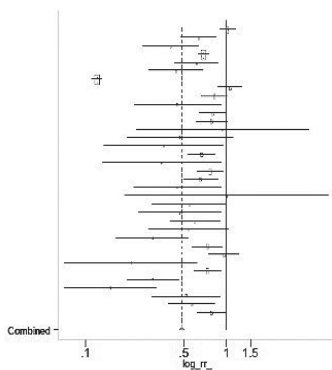
[27] Par exemple, dans le cas d'un projet de développement urbain, si une augmentation des prix du foncier est constatée, créant un effet d'éviction pour les résidents les plus pauvres, il est possible qu'une fraction plus importante de ménages (non aléatoire/non exogène au projet) ne puisse être réinterrogée dans la zone bénéficiaire que dans la zone contrôle, de sorte que l'impact estimé sur des variables de revenu s'en trouve biaisé à la hausse.

- Un risque également bien documenté (Ravallion, 2011), mais souvent peu discuté dans les papiers de restitution des résultats de ces travaux, est celui du biais d'expérimentation, mis en avant dès la fin des années 1920 aux États-Unis (effet Hawthorne constaté lorsque les résultats d'une expérience ne sont pas dus aux facteurs expérimentaux, mais au fait que les sujets ont conscience de participer à une expérience dans laquelle ils sont testés, ce qui se traduit généralement par une plus grande motivation ; effet John Henry, qui est semblable à l'effet Hawthorne, mais concerne le groupe de contrôle – par exemple, si celui-ci « ne veut pas perdre la face »). Ce biais est lié au fait que dans la majorité des expérimentations sociales, contrairement aux essais cliniques, les individus connaissent leur statut (bénéficiaire/non-bénéficiaire), de même que les personnes en charge de l'évaluation. L'on ne se situe donc pas dans une situation de « double aveugle ». Or, les quelques cas de travaux où cette précaution a pu être appliquée montrent que ce biais n'est potentiellement pas négligeable (voir graphiques 6 et 7 sur des évaluations d'impact relatives à l'amélioration de la qualité de l'eau).

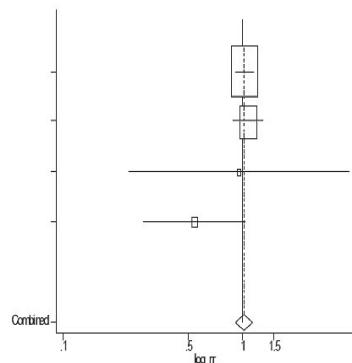
Graphiques 6 et 7

Un biais en chasse un autre : impacts de l'amélioration de la qualité de l'eau avec/sans double aveugle

L'ensemble des études mobilisées conclut que les interventions sur la qualité de l'eau améliorent fortement les indicateurs de santé infantile (en moyenne division par deux de l'incidence de la diarrhée)...



... Les quatre études menées en double aveugle concluent à un impact en moyenne non significatif.

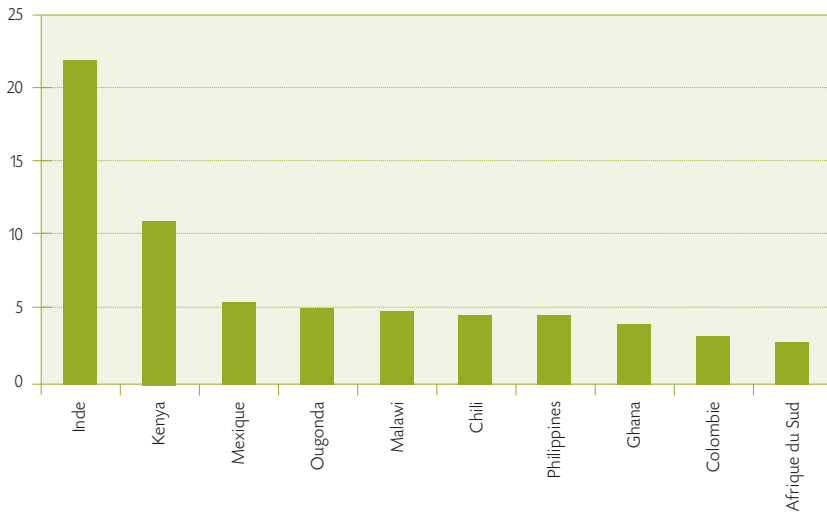


Source : Cairncross (2012).

- De vives discussions portent également sur les limites en termes de validité externe de la méthode RCT (encore que cette limite est commune – dans une certaine mesure – à l'ensemble des méthodes d'évaluations d'impact). Cette critique fait référence au caractère très particulier/contextuel des résultats obtenus, et aux difficultés de généralisation de ces résultats. Cela peut tenir à l'étroitesse de l'échantillon d'étude retenu, nécessaire pour s'assurer de la validité interne du protocole, mais parfois peu représentatif de la population cible (voir Bernard *et al.*, 2012 ; Parienté, 2008). Cela peut tenir également au contexte très spécifique de l'expérimentation (par exemple, territoire très particulier, ONG mettant en œuvre le programme très motivée, etc.). Sur ce dernier point, les revues systématiques de la littérature, qui se sont fortement développées parallèlement à l'expansion des évaluations d'impact, apportent des éléments de réponse (mais qui sont eux-mêmes contestés, cf. *Overseas Development Institute* [ODI], 2012). En outre, la relative concentration géographique des expérimentations ne peut que renforcer la crainte d'une validité externe limitée (cf. graphique 8). Enfin, peut-être encore plus fondamentalement, cela peut tenir à l'échelle du programme testé. Ainsi, les dispositifs évalués à une échelle locale peuvent avoir – lorsqu'ils sont passés à plus grande échelle – des effets systémiques/macroéconomiques par définition non mesurés par un « pilote ». Un exemple classique cité par Gurgand et Valdenaire (2012) est l'expérimentation de la réduction de la taille des classes aux États-Unis (expérimentation STAR). Cette expérimentation sur quelques dizaines de classes, dans lesquelles les élèves ont été affectés aléatoirement, a donné des résultats probants. Généralisée, elle implique cependant des recrutements massifs d'enseignants, comme ce fut le cas dans certains États américains, inspirés par les résultats de cette expérimentation. Or, certains craignent que le niveau de formation plus bas de ces nouveaux enseignants ne contrebalance l'effet de réduction de la taille des classes. De tels effets de bouclage négatifs sont également mis en évidence en France dans le cadre d'une des trop rares études tentant de mesurer ces effets (Crépon *et al.*, 2011) sur l'accompagnement renforcé des chômeurs. Une réflexion similaire est menée par Billé (2009) sur le « mythe de l'effet tâche d'huile ».

Graphique 8 Une nette concentration géographique des expérimentations

Répartition géographique des évaluations d'impact conduites par le J-Pal
(en % du total des évaluations d'impact conduites hors celles conduites aux États-Unis et dans l'UE)



Note de lecture : dix pays concentrent les 2/3 des expérimentations conduites par le J-Pal dans les pays en développement ; l'Inde et le Kenya représentent à eux seuls 1/3 de ces évaluations. L'analyse de la base d'évaluations d'impact de la Banque mondiale fait également ressortir ces deux pays parmi les trois principales zones d'études, après le Brésil (qui concentre à lui seul environ 8% des évaluations d'impact de la Banque sur la période 2005-2010). L'IEG (2012) note également qu'en 2009 – 2010, un quart des opérations financées par la Banque mondiale au Brésil a fait l'objet d'une évaluation d'impact.

Source : J-Pal, calculs de l'auteur.

- Enfin, la question de la validité « conceptuelle » (*construct validity*) des expérimentations, c'est-à-dire leur capacité – au travers des variables de mesure choisies – à apprécier de façon pertinente les évolutions en jeu, est également débattue (Johnson, 2013), de même que celle de la qualité des données sur lesquelles elles reposent. À ce titre, les travaux de réplcation de Palmer-Jones (2013) sont tout à fait intéressants, qui montrent que la robustesse des résultats établis repose *in fine* tout autant sur la fiabilité du protocole, que sur la qualité des données collectées (Ravallion, 2012). De façon liée, Gibson *et al.*, (2013) mettent en évidence des erreurs de mesure de la consommation des ménages issues des enquêtes ménages, qui seraient susceptibles de biaiser les résultats des évaluations d'impact.

2.2.2. ...à un retour à une plus grande pluralité de méthodes

Les critiques fortes portées à l'encontre de l'expérimentation aléatoire contrôlée finissent par se répercuter non seulement sur le positionnement de certains bailleurs, mais aussi sur celui des fonds multilatéraux comme le 3IE et le SIEF, en faveur d'un consensus pragmatique plus équilibré. Celui-ci s'est manifesté clairement lors de la réunion du NONIE en 2012 qui a évoqué la nécessité de passer du « *gold standard* » à de « *good standards* », mais aussi par un papier du DFID la même année (Stern *et al*, 2012), invitant les bailleurs à élargir la palette des méthodes utilisées pour évaluer l'impact des projets de développement. De façon notable, le 3IE fait évoluer dans ses nouveaux appels à projet les critères de sélection des évaluations, afin de ne pas exclure, comme par le passé, les évaluations non expérimentales. Cet assouplissement méthodologique est également constaté dans le SIEF II. Le 3IE joue également un rôle actif dans l'évolution des standards concernant les revues systématiques de la littérature vers une plus grande mixité d'approches – pour inclure des études reposant sur des méthodes non expérimentales et qualitatives (cf. encadré 6), ainsi que la littérature grise (AusAid, 2011 ; Dzakupasu *et al*, 2013 constitue un exemple d'une revue élargie).

Encadré 6 Évaluations d'impact : vers une plus grande mixité de méthodes

Au-delà des évaluations d'impact expérimentales, il existe trois autres grands types de méthodes dites quasi-expérimentales. Le principal avantage de ces méthodes, d'essence économétrique, est leur plus grande faisabilité pratique (puisque l'on a vu que l'expérimentation, et en particulier le principe aléatoire, n'était applicable que dans un nombre réduit de cas). Toutefois, l'inconvénient de ces méthodes est de reposer sur un certain nombre d'hypothèses dites « identifiantes », et de produire des résultats, dans une certaine mesure, sensibles aux spécifications économétriques adoptées (Givord, 2010).

La méthode de l'appariement (*matching*) est parfois présentée comme une sorte de second best par rapport à la méthode RCT. Elle consiste à appairer un à un les individus (ou unités statistiques) d'un groupe de non-bénéficiaires aux individus bénéficiaires sur la base de caractéristiques observables (par exemple, l'âge, le revenu, etc.). L'avantage de cette méthode est qu'une fois les deux groupes constitués, les estimations d'impact sont réalisées en calculant « simplement » la différence entre la moyenne des variables d'intérêt, sans reposer sur un modèle sous-jacent linéaire. Néanmoins, cette méthode repose sur l'hypothèse que les caractéristiques non observables des individus béné-

...

...

ficiaires et non-bénéficiaires sont parfaitement corrélées à leurs caractéristiques observables (sinon, il existera un biais dans les estimations). Elle nécessite en outre en pratique de disposer d'un nombre important de données et d'un « support commun » suffisamment large (i.e. d'un nombre d'individus comparables suffisamment grand). La méthode de « différence-de-différence » (*difference-in-difference*) consiste à comparer la situation des individus bénéficiaires à celle des non-bénéficiaires avant et après l'intervention (dont on cherche à mesurer l'impact). En combinant une méthode « avec/sans » et une méthode « avant/après », elle vise donc à contrôler à la fois les biais temporel et individuel. Cette méthode, qui repose en pratique sur des régressions linéaires, suppose que la différence observée avant l'intervention entre groupe de bénéficiaires et groupe de contrôle, constitué de façon ad hoc, soit constante dans le temps. Ainsi, si elle est conceptuellement simple, cette méthode présente des limites importantes. La technique des variables instrumentales consiste enfin à utiliser une variable corrélée avec l'intervention dont on cherche à mesurer l'impact, mais qui n'a pas d'impact direct sur les résultats de cette mesure (exogène par rapport à la variable d'intérêt). La qualité des résultats obtenus avec cette méthode va toutefois dépendre de la validité de l'instrument choisi et, en pratique, il peut être difficile de trouver une variable adaptée. À noter que ces différentes méthodes peuvent être dans une certaine mesure combinées. Enfin, il existe également des méthodes d'évaluations d'impact plus qualitatives, notamment la méthode des contributions (SFE, 2010).

Dans ce contexte, la recherche de la *mesure* des impacts précisément attribuables à un projet semble devenir moins centrale que celle d'une appréciation plus lâche de ses impacts (objectif d'attribution sans nécessairement de mesure précise de la taille des effets ; Stern *et al.*, 2012). Certains chercheurs évoquent quant à eux la recherche de l'établissement d'une « plausibilité » des liens de cause à effet, en mobilisant différentes approches, plutôt que de la recherche de l'établissement strict d'une « causalité » au sens statistique du terme (Cairncross, 2012). D'autres évoquent la nécessité de se diriger vers plus de pragmatisme dans les méthodes utilisées (Ridde et Haddad, 2013). Des prises de position sur la nécessité de « faire le deuil de l'attribution » pour se contenter d'un objectif de *contribution* à des résultats/des impacts se font également jour (voir séminaires AFD – F3E, séminaire SFE en 2013)^[28]. Une attention croissante dans les évaluations d'impact est par ailleurs portée à l'étude de la « chaîne de l'impact » ou à la « théorie du programme », constituant l'amorce d'une réconciliation entre évaluations « standards », axées sur les processus de mise

[28] Au travers de la méthode d'analyse des contributions par exemple (voir SFE, 2010).

en œuvre et d'atteinte des premiers résultats, et évaluations d'impact (White, 2009 ; voir également encadré 7 sur quelques travaux de la coopération néerlandaise). En termes méthodologiques, le mixage des méthodes (quantitatives, qualitatives) et l'adéquation de la méthode à l'objet sont encouragés (dès 2007 par la Société européenne d'évaluation^[29] et dès 2009 par le NONIE). Le 3IE invite désormais clairement à une telle mixité (« *Quality impact evaluations require mixed methods; that is a combination of both quantitative and qualitative methods* »)^[30] et marque, avec la création du *Policy Window* (dans lequel des projets soumis par leurs opérateurs – gouvernements/ONG – sont préalablement sélectionnés en fonction de leur pertinence, puis proposés à des équipes de chercheurs en charge de définir un protocole), la volonté de ne pas « être guidé » par les méthodes^[31]. *De facto*, les méthodes non expérimentales et qualitatives semblent ainsi faire l'objet d'une forme de réhabilitation (« *It's a good time to be a qualitative researcher in impact evaluation...* », Camfield, 2013). Des méthodes alternatives d'évaluations d'impact concernant les projets « à petit n » émergent (White et Phillips, 2012 ; McKenzie, 2011a pour les interventions portant sur des entreprises). Banerjee et Duflo (2009), face aux critiques grandissantes sur le « tout-expérimental », invitent même à une utilisation complémentaire des méthodes expérimentale et quasi-expérimentale pour étudier les effets d'équilibre général des programmes. Enfin, la création du centre de recherche *Centre of Development Impact* (CDI) en mars 2013, initiative conjointe de l'*Institute of Development Studies* (IDS) et de l'ITAD^[32], montre la volonté de proposer une offre structurée s'appuyant sur l'ensemble des méthodes des sciences sociales pour évaluer les impacts des interventions de développement.

[29] In September 2007, the European Evaluation Society (EES) felt the need to issue a statement about "the importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and in development interventions", pointing out that "the EES supports multi-method approaches to impact evaluation and does not consider any single method such as RCT as first choice or as the 'gold standard'."

[30] Voir dans « *3ie Impact Evaluation Practice: a guide for grantees* » (site du 3IE).

[31] Même si certains chercheurs pointent qu'en pratique, les approches mixtes restent souvent dominées par le volet quantitatif.

[32] Bureau de consultants en économie du développement basé à Brighton (Royaume-Uni).

Encadré 7 *Évaluations d'impact versus autres types d'évaluation : un – désormais – faux débat ?* *Le cas de la coopération néerlandaise*

L'évaluation d'impact comme sous-composante d'une évaluation plus large : le cas de l'évaluation des aides budgétaires...

La coopération néerlandaise a mené plusieurs travaux d'évaluation de ses programmes d'aide budgétaire (sectorielle ou globale), incorporant une composante évaluation d'impact. Par exemple, dans le cadre de l'évaluation d'un aide budgétaire globale apportée à la Zambie (menée de façon conjointe avec les coopérations allemande et suédoise), elle a incorporé des évaluations d'impact portant sur les cinq secteurs d'activité auxquels l'aide budgétaire a principalement contribué. Comme il est indiqué dans IOB (2011a), dans ce cas, il ne s'agit pas tant de mesurer l'impact de l'aide-budgétaire, que celui des politiques mises en place par le gouvernement zambien, auxquelles l'aide-budgétaire a permis de contribuer. Ces études ont mobilisé des approches mixtes (quasi-expérimentales et qualitatives) et principalement des données existantes. En outre, un modèle d'équilibre général a été parallèlement utilisé afin d'estimer les effets macroéconomiques des flux d'aide en Zambie. Ces travaux sont intéressants à double titre : d'une part, ils illustrent la complémentarité possible entre évaluation d'impact et évaluation plus classique de programmes ; d'autre part, ils montrent les possibilités d'utilisation des évaluations d'impact pour tenter d'apprécier les impacts (ou les contributions aux impacts) d'instruments plus larges et plus complexes que l'aide projet classique.

...et de projets dans le secteur de l'eau

La coopération néerlandaise a conduit plusieurs évaluations d'impact dans le domaine de l'eau et de l'assainissement, dans lesquelles, outre l'analyse d'impact *stricto sensu*, d'autres aspects d'une évaluation plus classique (au sens des standards de l'OCDE) ont été étudiés de façon approfondie, comme le contexte et la pertinence des interventions, leur mise en œuvre ou encore leur durabilité. C'est le cas en particulier d'une étude menée avec la coopération allemande au Bénin (IOB, 2011b). Ces exemples sont également intéressants dans le sens où ils illustrent le fait qu'opposer évaluation d'impact et évaluation classique ou « de processus », comme ce qui est souvent fait, n'est peut-être plus totalement pertinent.

Pour conclure cette partie, l'on peut noter qu'au-delà des débats intenses autour des vertus de la méthode expérimentale, l'une des externalités très positive de « l'apparition » de cette méthode dans le domaine de l'évaluation aura été de rehausser les standards de qualité globaux de l'évaluation dans le domaine du développement (Picciotto, 2012 ; voir l'exemple de l'évaluation des aides budgétaires^[33]), tout en créant un renouveau de la recherche en économie du développement (Rosenzweig, 2012).

2.3. Une clarification des attentes et des objectifs fixés aux évaluations d'impact

L'objectif initial des évaluations d'impact est très ambitieux : les évaluations d'impact visent à accroître la prise de décision – à la fois chez les bailleurs et les décideurs politiques des pays du Sud – basée sur la preuve (*evidence-based policy making*). Il va donc au-delà du rôle habituel dévolu à la recherche, qui consiste à accroître la somme des connaissances (nouvelles)^[34], pour viser à influencer directement ou indirectement (dans le sens d'une amélioration) sur la formation des politiques publiques (Center for Global Development [CGD], 2006^[35]). C'est au demeurant cette finalité qui est mise en avant pour justifier le coût des évaluations d'impact (« [...] *the cost of conducting an impact evaluation must be compared to the opportunity costs of not conducting a rigorous evaluation and thus potentially running an ineffective program* », Gertler et al, 2011). De façon incidente, en apportant la preuve de « ce qui marche » et de « ce qui ne marche pas », les évaluations d'impact doivent également permettre de répondre aux besoins de redevabilité des bailleurs (c'est-à-dire de rendre compte de l'efficacité de leurs interventions, en grande partie financées sur fonds publics). En revanche, la première génération d'évaluations d'impact s'intéresse relativement peu aux chaînes de cause à effet permettant d'expliquer *comment* les impacts mesurés ont été produits (notamment en raison de contraintes méthodologiques et parce que l'étude plus factuelle des réalisations et premiers résultats générés par les projets est renvoyée aux évaluations qualifiées de processus ; cf. J-Pal, 2011a).

[33] Issue favorable après une période incertaine où il était à craindre qu'au contraire, l'activité d'évaluations d'impact en général pâtisse du courant « anti-expérimentation ».

[34] Sans présumer de l'utilisation qui sera faite des résultats (position défendue en France par le LIEPP – Laboratoire interdisciplinaire d'évaluation des politiques publiques).

[35] "Addressing this gap, and systematically building evidence about what works in social development, would make it possible to improve the effectiveness of domestic spending and development assistance by bringing vital knowledge into the service of policymaking and program design".

Tableau 6 *Rôle des évaluations d'impact :
que reste-t-il des objectifs initiaux ?*

Rôle des évaluations d'impact	Type d'utilisation	Attentes initiales (CGD, 2006)	Attentes revues
Objectif informatif : prendre des décisions informées (<i>evidence-based decisions</i>)	Utilisation directe (maintien/arrêt/ passage à l'échelle d'une intervention)	Oui (justifie le coût des évaluations d'impact)	Pas toujours, pas souhaitable
Objectif accumulatif : contribuer à l'accumulation de connaissances sur « ce qui marche » et « ce qui ne marche pas » et <i>in fine</i> sur les ressorts du développement	Utilisation indirecte/ stratégique (alimentation des stratégies des décideurs politiques/ des bailleurs/réplication)	Oui	Oui mais (problème de la « théorie du programme » ; problème de la contextualisation)
Objectif sommatif : démontrer l'efficacité d'une intervention	Redevabilité/plaidoyer	Oui	Pas toujours (Bernard <i>et al.</i> , 2012) ^[36]
Objectif formatif : comprendre les chaînes de cause à effet	Apprentissage/utilisation directe (amélioration des interventions actuelles/ futures)	Pas central	Oui de plus en plus (approches basées sur la théorie, méthodes mixtes, expérience de DIME)

2.3.1. Prendre des décisions fondées sur la « preuve » ? (Utilisation directe)

Si, dans certains cas, une utilisation directe des résultats des évaluations d'impact dans la prise de décision a bien été constatée (exemples emblématiques de Progres, des vermifuges au Kenya, etc.), dans de nombreux autres cas, il ne semble pas que les preuves solides (*hard evidence*) apportées par ces travaux aient été davantage valorisées et intégrées dans la prise de décision que les preuves lâches

[36] Le rôle des évaluations d'impact en termes de redevabilité, largement discuté par Bernard *et al.*, (2012), n'est pas développé dans ce papier. À l'issue des travaux menés par l'AFD, la division de l'Évaluation est d'avis que l'objectif de redevabilité vis-à-vis des contribuables pourrait être mieux atteint par des outils d'évaluation plus légers (que les évaluations d'impact), qui se concentrent sur la mise en œuvre des projets (bonne utilisation des fonds, efficacité en termes de coût et de délai, régularité des passations de marchés, réalisations effectives, etc.), tandis que l'objectif de redevabilité vis-à-vis des partenaires du Sud pourrait s'attacher à analyser la pertinence de l'intervention par rapport au cadre stratégique défini par le pays et à ses besoins et la valeur ajoutée du bailleur.

(voire anecdotiques) du passé (Dercon, 2013). Ainsi, la Banque mondiale (IEG, 2012), qui constitue le bailleur au sein duquel la pratique d'évaluations d'impact est la plus institutionnalisée, constate que seules 22 % à 33 % de ces études (selon les sources) ont eu des répercussions opérationnelles directes (pour poursuivre les projets, favoriser leur passage à l'échelle, au contraire les réduire ou les abandonner, ou enfin améliorer la mise en œuvre de secondes phases des projets). L'IEG identifie plusieurs raisons expliquant cette utilisation directe modeste : dans certains cas, le mauvais *timing* en termes de production des résultats par rapport au temps de la décision ; l'implication insuffisante des différentes parties prenantes (cf. section 2-3-3) ; le manque de diffusion des résultats ; la pertinence pas toujours avérée des questions évaluatives traitées par rapport aux objectifs/questions opérationnelles posées par le projet ; enfin, le manque de culture de l'évaluation. Pour illustration, dans le cas de l'expérience de l'AFD, il est clair que les chefs de projet ne valorisent pas forcément davantage la preuve apportée par les évaluations d'impact (dont ils ne perçoivent pas nécessairement la supériorité technique ; voir également IEG, *ibid.*), que celle issue de travaux reposant sur des méthodologies plus légères. C'est bien la capacité des études de type recherche-action à répondre aux questions opérationnelles qu'ils se posent, et au bon moment, qui semble déterminante (cf. encadré 8). De façon plus générale, la volonté même des services opérationnels des agences de développement d'utiliser les résultats de la recherche, et la valeur qu'ils leur accordent semblent faire encore en partie défaut : ainsi, une étude récente menée par le réseau *Poverty Reduction, Equity and Growth Network* (PEGNet) (Kleemann et Böhme 2013) auprès de praticiens du développement en Allemagne indique que 40 % d'entre eux se disent « avec satisfaction » non informés des résultats de la recherche (*happily uninformed*). Weyrauch et Langou (2011) reconnaissent également l'importance des difficultés qui jalonnent le chemin allant de l'établissement des résultats à l'influence opérationnelle/politique.

Encadré 8 Études sommatives versus études formatives : retours d'expérience de l'AFD, du MCC et de la JICA

Dans le domaine de l'éducation, l'évaluation sommative se définit comme une évaluation intervenant au terme d'un processus d'apprentissage ou de formation afin de *mesurer* les acquis de l'élève ou de l'étudiant. L'évaluation formative quant à elle est une évaluation qui a pour fonction d'améliorer *l'apprentissage* en cours en détectant les difficultés de l'élève afin de lui venir en aide, en modifiant la situation d'apprentissage ou le rythme de cette progression, pour apporter (s'il y a lieu) des améliorations ou des correctifs appropriés. De façon plus générale, l'évaluation sommative a pour fonction de *rendre compte* ou de certifier/démontrer des résultats, tandis que l'évaluation formative vise plutôt à se rendre compte/à fournir un diagnostic permettant d'améliorer le processus de mise en œuvre d'une intervention/d'une politique (cf. tableau 7). Bien sûr, malgré cette distinction, ces deux types d'évaluation ne sont pas sans liens, et certains travaux visent à intégrer ces fonctions redevabilité et apprentissage. En matière d'évaluations d'impact, les expériences respectives de l'AFD, du MCC et de la JICA permettent de fournir des illustrations sur les tensions et/ou les frustrations, qui peuvent toutefois exister lorsque ces travaux d'évaluation poursuivent des objectifs multiples.

Tableau 7 Évaluation sommative et évaluation formative : quelques éléments distinctifs

	Évaluation sommative	Évaluation formative
Fonction	Redevabilité (rendre compte)	Diagnostic (comprendre)
Comment ?		
Évaluateur	Externe (pour crédibilité)	Interne (pour appropriation)
Type d'outils	Surtout quantitatifs	Surtout qualitatifs
Méthode	Normée	Adaptative
Échantillon	Représentatif	Choisi à dessein

Source : inspirée de Cairncross (2012).

Dans le cas de l'AFD, l'évaluation d'impact de la micro-assurance de santé SKY a permis de fournir certaines preuves solides sur les résultats/les impacts de la micro-assurance de santé en milieu rural au Cambodge : ainsi, il est établi (Levine *et al.*, 2012) que celle-ci a des impacts économiques significatifs et importants sur les ménages

...

...

bénéficiaires, et qu'elle permet efficacement de réorienter l'utilisation des services de santé du secteur privé vers le secteur public (jugé de meilleure qualité) ; en revanche, elle n'aurait pas d'impact sanitaire (tout au moins, sur l'horizon court de l'étude, un tel impact ne peut être démontré). Toutefois, les vertus de cette évaluation en termes d'apprentissage sont contestées par certaines parties prenantes de l'étude (Quentin et Guérin, 2013). Pour illustration : certaines questions jugées clés en termes d'*apprentissage* par l'opérateur du dispositif, comme les effets de la micro-assurance de santé sur le fonctionnement du système de santé public, ou encore l'hétérogénéité des impacts, sont peu étudiés dans la mesure où le cadre méthodologique adopté ne le permet pas. D'autres résultats, malgré leur caractère novateur d'un point de vue académique (étude de la sélection adverse, par exemple), sont considérés comme peu pertinents dans le contexte d'un pays en développement, confronté à un problème de sous-consommation des services de santé. Au final, l'on voit que, dans ce cas, les objectifs multiples assignés à ces travaux (démontrer des impacts ; comprendre le chemin de l'impact ; etc.), mêlant fonctions sommative et formative, n'ont pas pu tous être atteints de façon égale. Un contre-exemple à cette étude est celle, également soutenue par l'AFD à peu près au même moment, sur l'évaluation des politiques publiques dites de « gratuité » des soins de santé dans trois pays du Sahel (Burkina Faso, Mali et Niger ; cf. Oliver de Sardan et Ridde, 2012). Ces travaux, basés sur des méthodes non expérimentales, ont bénéficié d'un écho très important tant au niveau de l'AFD, qu'au sein des pays concernés. La pertinence des questions traitées et la souplesse de la méthode adoptée (utilisation de données existantes, volet qualitatif) semblent avoir été les clés de l'influence importante de cette étude, illustrant le fait que les services opérationnels (et l'ensemble des parties prenantes) valorisent davantage ces éléments que l'utilisation de méthodes *a priori* plus rigoureuses (comme la méthode expérimentale), dès lors que ces travaux ont fourni des éléments de preuve jugés suffisamment crédibles.

L'exemple du MCC est également intéressant à ce titre. Ce bailleur a récemment publié un bilan des cinq premières évaluations d'impact expérimentales, qu'il a réalisées dans le domaine de l'agriculture (MCC, 2012). Ces travaux ont porté sur le même type d'intervention (formation aux agriculteurs) dans 5 pays différents. Si les études menées permettent de démontrer que cette intervention a généré une hausse des revenus agricoles dans deux pays (et une partie d'un troisième pays) et, dans tous les cas, n'a pas eu d'impact (sur l'horizon de mesure) sur les revenus des ménages, elles apparaissent relativement limitées d'un point de vue formatif : de l'avis du MCC, des analyses complémentaires seront nécessaires pour comprendre l'hétérogénéité des impacts ; d'un point de vue opérationnel, les conclusions (importance du timing dans la mise en œuvre des différentes composantes des projets, importance du format des formations, etc.) paraissent limitées par rapport aux efforts déployés.

...

...

Enfin, la JICA a pour sa part conduit 4 évaluations d'impact dans le domaine des infrastructures d'irrigation, en faisant appel à des méthodes quasi-expérimentales (JICA, 2010). Ces études permettent d'établir, dans certains cas, que ces infrastructures ont généré un impact positif sur la production agricole des zones concernées, et apportent des éléments intéressants en termes de mesure de la rentabilité de ces infrastructures. Une certaine hétérogénéité des impacts est constatée, en lien avec la qualité de la distribution d'eau en amont et en aval. Là encore, si la fonction démonstrative de cette étude est forte, sa fonction formative peut être jugée insuffisante (la principale conclusion opérationnelle étant que la qualité de la distribution d'eau est un facteur critique pour que la mise en place d'infrastructures d'irrigation permette une augmentation significative de la production agricole).

De façon plus fondamentale, l'on peut se demander s'il est souhaitable de prôner *dans tous les cas* un modèle d'utilisation déterministe des résultats des travaux d'évaluation d'impact. L'on peut ainsi citer plusieurs cas dans lesquels une utilisation naïve des résultats de ces travaux est à proscrire :

- Ainsi, dans le cas des projets d'adduction en eau potable, il est désormais établi que les impacts sanitaires sont souvent plus limités que ce qui est clamé dans les projets de développement (Snilstveit et Waddington, 2009 ; Caircross et Valdmanis, 2006 notamment) ; en revanche, leurs impacts socioéconomiques ont été globalement peu étudiés, mais sont potentiellement plus significatifs (réduction du temps de la corvée d'eau, des dépenses liées, etc.). Le même constat est fait par Dercon (2013) dans le domaine de la micro-assurance, pour qui les éléments de preuve disponibles seront toujours incomplets par rapport à l'immensité de la « nature ». Il se demande s'il faut pour autant attendre ces preuves pour prendre des décisions.
- En outre, il a déjà été mentionné que les évaluations d'impact peinent le plus souvent à mesurer des effets à moyen et long termes (dans le cas de la Banque mondiale, l'IEG (2012) note que seulement 12 % des évaluations d'impact conduites essaient de mesurer des effets sur des variables à moyen/long termes). Ainsi, un projet jugé négativement sur la base des impacts produits au bout de un à deux ans pourra en réalité se révéler un succès si une évaluation d'impact pouvait être menée au bout par exemple de cinq ans. Cela peut se révéler particulièrement important dans le cas de produits/services nouveaux dont l'adoption par les populations sera lente (cas de l'étude *Al Amana* financée par l'AFD ; Bernard *et al.*, 2012). On comprend bien qu'il est, dans ce

- Dans d'autres cas, un échantillon d'études trop petit par exemple pourra se traduire par une difficulté à détecter des impacts significatifs, mais ne permettra pas de préjuger de l'effectivité ou non de ces impacts (« *Absence of evidence is not evidence of absence* », Levine et al., 2012).
- Par ailleurs, malgré l'objectivité des résultats issus de ces travaux, une part de subjectivité reste présente dans leur interprétation. Ainsi, dans le cas de l'évaluation de la microassurance de santé SKY au Cambodge menée par l'AFD (cf. encadré 8), deux interprétations presque opposées des résultats sont possibles selon la « fonction de préférence » des décideurs : sur la base des impacts économiques obtenus et de la typologie des adhérents (généralement, les personnes les plus malades), le programme peut être considéré comme un succès ; toutefois, sur la base des indicateurs de santé (utilisation des services de santé, état de santé) et du taux d'adhésion (très faible hors subventions), le programme peut en revanche être jugé négativement. C'est donc un diagnostic nuancé, faisant appel à d'autres travaux plus qualitatifs, qui devra en réalité être fait et mobilisé pour toute prise de décision (Gurgand et Valdenaire, 2012)^[37].
- Le cas de l'étude Progres sur les transferts monétaires conditionnels est également à ce titre tout à fait intéressant : malgré le succès incontesté de cette étude (notamment en termes d'utilisation), des débats vifs sont apparus ultérieurement sur l'interprétation des résultats et en particulier sur le rôle de la conditionnalité initialement mise en avant (Ozler, 2013).
- Enfin, comme évoqué, les quelques études qui explorent les effets d'équilibre général des interventions montrent l'écart important qu'il peut y avoir entre impacts locaux et impacts nets globaux, et par conséquent les difficultés de fonder les décisions directes de passage à l'échelle d'une intervention uniquement sur la base des résultats d'une évaluation d'impact de taille réduite. Pour exemple, Finkelstein (2005) établit que l'impact sur les dépenses de santé aux Etats-Unis de l'assurance Medicare était au moins quatre fois supérieur à ce qui avait été estimé par l'étude expérimentale RAND menée à petite échelle dans les années 1990.

[37] « Il faut absolument renoncer à un usage naïf de l'évaluation : elle ne donne pas des réponses toutes faites, elle ne promeut pas des dispositifs « clés en main » que l'on peut dupliquer à la chaîne [...]. Au contraire, elle donne des éléments qui [...], ne constituent qu'une pièce d'un mécano [...] que l'on devra inclure dans un ensemble, qui reste à penser, avec ce que cela comporte de risque et de choix politique ».

2.3.2. *Bâtir des stratégies fondées sur la « preuve » ? (Utilisation indirecte)*

L'utilisation indirecte des résultats des évaluations d'impact dans la définition des stratégies des bailleurs et des décideurs politiques est sans doute plus importante, mais aussi plus difficile à apprécier. La Banque mondiale (IEG, 2012) note ainsi que dans les domaines ayant bénéficié d'un nombre important d'évaluations d'impact de haute qualité, il semble bien qu'elles aient influencé la stratégie/le type d'interventions promues par la Banque (comme dans le secteur de l'éducation et de la protection sociale). Deaton (2010) pointe néanmoins le risque que certaines évaluations d'impact, peu regardantes sur le processus de production des résultats obtenus (ou sur la « théorie des programmes »), se révèlent peu utiles pour l'accumulation de connaissances sur les ressorts du développement. De façon liée, pour l'ODI (2012), les leçons qui peuvent être tirées des revues systématiques de la littérature, censées mettre en évidence des résultats « généralisables » sur « ce qui marche » et « ce qui ne marche pas », seraient peu utilisables pour les décideurs politiques dans une optique de « répliquabilité » des interventions étudiées. Elles seraient davantage d'ordre académique. En effet, dans le domaine des sciences sociales, le contexte s'avère clé dans l'atteinte des résultats, et identifier les raisons du succès ou non des interventions (aspects insuffisamment traités dans ces revues) est tout aussi important que d'établir si elles ont généré ou non les résultats espérés.

2.3.3. *Quelle demande et quelle utilisation des évaluations d'impact par les pays partenaires ?*

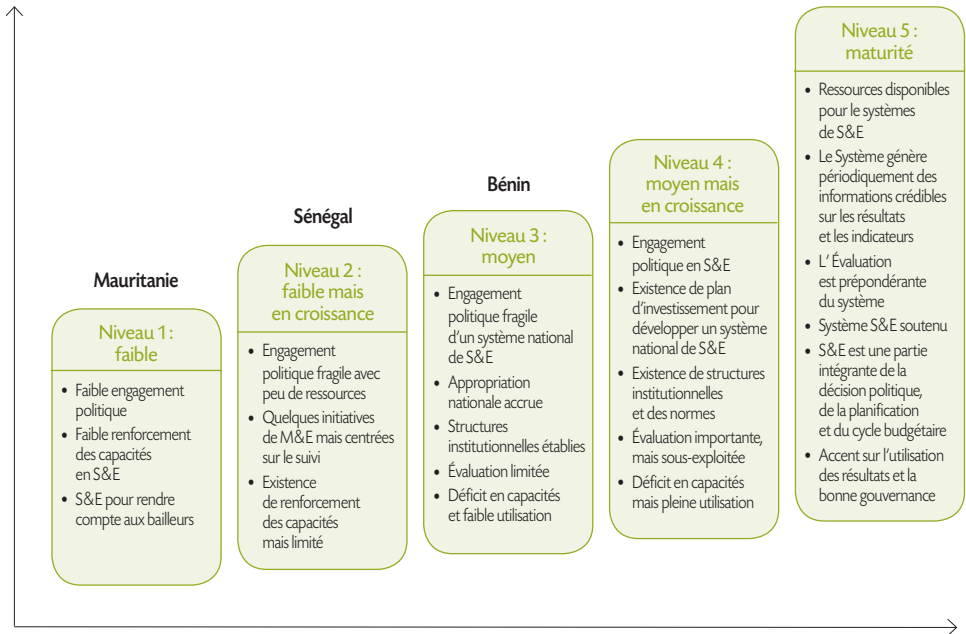
L'IEG (*ibid.*) montre que le facteur déterminant dans l'utilisation et l'influence de ces travaux est l'implication des pays partenaires. Ainsi, les évaluations d'impact conduites par la Banque mondiale, dans lesquelles les gouvernements des pays partenaires ont été impliqués dès la phase de conception, ont été utilisées dans le cadre de phases ultérieures du projet ou d'un nouveau projet similaire dans près de 60 % des cas ; elles ont influencé le dialogue politique de la Banque avec ces gouvernements dans près de 80 % des cas. L'IEG note toutefois que la proportion d'évaluations d'impact conduites à la Banque mondiale à l'initiative des pays partenaires reste faible (environ 26 % du total), celles-ci étant le plus souvent initiées par les équipes opérationnelles ou de chercheurs/évaluateurs de la Banque. De façon plus générale, il semble qu'il y ait une très forte hétérogénéité dans la demande d'évaluations (d'impact) de la part des pays partenaires et dans l'utilisation qui en est faite. L'environnement politique (notamment son caractère démocratique et le

rôle moteur d'une institution – comme le Parlement, le ministère des Finances ou la Présidence), ainsi que la place accordée aux systèmes de suivi-évaluation dans les pays sont reconnus comme des facteurs déterminants dans la demande et l'appropriation qui est faite des évaluations (d'impact) par les pays partenaires (IEG World Bank [2009] ; Gaarder et Briceno [2010]). Ainsi, il n'est pas surprenant que des pays considérés comme leaders en matière de système de suivi-évaluation (comme par exemple, en Amérique latine, le Chili, la Colombie et le Mexique) semblent plus demandeurs et utilisateurs de ce type d'études.

En Afrique subsaharienne, deux études menées par le *Center for Learning, Evaluation and Results* (CLEAR) (en 2012 sur neuf pays et en 2013 sur la Mauritanie, le Sénégal et le Bénin) montrent que, à l'exception peut-être du Bénin, la plupart des pays se situent à un stade de « mise en place » de leur système d'évaluation (cf. schéma 3 pour la Mauritanie, le Sénégal et le Bénin). Dans certains pays, les évaluations conduites sont surtout réalisées pour rendre compte aux bailleurs (Ouganda, Sénégal, Mauritanie, par exemple). L'étude publiée en 2013 identifie toutefois, au sein des trois pays étudiés (Mauritanie, Sénégal et Bénin), une demande d'appui technique et de formation aux évaluations d'impact émanant de plusieurs directions de ministères et institutions publiques. De façon intéressante, l'étude de CLEAR de 2012 montre en outre qu'en dépit des 36 évaluations d'impact qui ont été conduites au Kenya par le J-Pal, ainsi que d'autres évaluations portées par les bailleurs (comme l'*Alliance for a Green Revolution in Africa*), le système d'évaluation de ce pays reste peu développé, ces travaux ayant été conduits de façon largement externe au gouvernement. Pour illustration, l'agence leader en matière de suivi-évaluation (*Monitoring and Evaluation Directorate* – MED), créée en 2003 et logée au sein du ministère de la Planification et du Développement national, n'aurait pas exercé d'influence sur ces évaluations (d'impact), que cela soit en termes de pilotage, de conception ou d'utilisation. Pour autant, cela ne signifie pas que, parmi les nombreuses évaluations d'impact conduites au Kenya, certaines n'aient pas influé sur les politiques publiques mises en place, comme l'illustre l'exemple célèbre de l'évaluation des comprimés contre les vers intestinaux (cf. encadré 9).

Plus récemment, des initiatives ont été prises afin de promouvoir la demande d'évaluations d'impact de la part des pays partenaires, à la fois au travers du 3IE mais aussi via la création de CLEAR, cofinancé par le DFID et la Banque mondiale notamment.

Schéma 3 État actuel du système de suivi-évaluation dans trois pays d'Afrique francophone



Source : CLEAR (2013)

Encadré 9 *Deux exemples d'évaluations d'impact utilisées par des gouvernements des pays partenaires*

L'une des évaluations d'impact les plus connues, celle relative aux comprimés pour lutter contre les vers intestinaux (et son efficacité relative très forte sur la fréquentation scolaire), a conduit le Premier ministre kenyan à mettre en place en 2009 un programme national, au niveau des écoles, de distribution de ces comprimés, géré par le ministère de l'Éducation. Selon le J-Pal, 3,65 millions d'enfants ont déjà bénéficié de ce programme, et une nouvelle campagne a été menée en 2012. Des gouvernements d'autres pays (Inde, Ethiopie, Gambie) ont également adopté des programmes similaires.

Un autre exemple concerne également le secteur de l'éducation, cette fois aux Philippines : en effet, la JICA a conduit en 2012 une évaluation d'impact, basée sur une méthode quasi-expérimentale, d'un programme visant à améliorer l'offre éducative dans les zones les plus pauvres des Philippines. Cette étude a permis de mettre en évidence des impacts positifs très marqués à la fois en termes de niveau scolaire des enfants et d'allongement de la durée des études. Les effets ressortent particulièrement marqués pour les femmes (hausse des inscriptions à l'université ; à long terme, hausse des revenus obtenus sur le marché du travail). Au final, le taux de rendement de ce programme a été estimé à plus de 15 %. Les résultats de ces travaux présentés au ministère de l'Éducation des Philippines devraient être directement utilisés dans le cadre d'une prochaine réforme du système éducatif, selon ce ministère (JICA, 2012).

2.3.4. Quel rôle assigner aux évaluations d'impact dans le contexte du développement ? Une solution à « deux angles »

Un usage plus sélectif des évaluations d'impact (en particulier expérimentales) semble désormais prôné, qui peut être résumé en deux types d'utilisation (cf. tableau 8) :

- d'une part, les évaluations d'impact peuvent être utilisées comme un outil de production de connaissances empiriques ;
- d'autre part, elles peuvent constituer un outil très opérationnel, d'optimisation de la mise en œuvre des projets.

Tableau 8 *Vers une utilisation mieux ciblée des évaluations d'impact*

Type d'utilisation	Utilisation précise	Que fait-on ?
Outil de production de connaissances empiriques	Mesure de l'impact de programmes nouveaux / innovants	Une évaluation d'impact « standard » : mesure de l'impact de ces programmes
	Mesure d'impacts sur des liens de cause à effet peu explorés (à la fois dans la littérature quantitative et qualitative)	Une évaluation d'impact « standard » ; par exemple, on mesure l'impact de l'accès à l'eau sur les dépenses des ménages (lien peu étudié empiriquement)
	Analyse des facteurs de réussite / d'échec des projets	Une évaluation d'impact davantage centrée sur la chaîne de l'impact et l'hétérogénéité (du contexte, de la mise en œuvre, des bénéficiaires), que sur la mesure statistique des impacts, avec un protocole mixte privilégiant le volet qualitatif
	Tests empiriques d'hypothèses théoriques (qui sous-tendent les projets de développement)	Une étude utilisant le programme comme véhicule au travers duquel on teste des comportements / des hypothèses
	Estimations utilisées pour le paramétrage de modèles macroéconomiques visant à simuler les effets globaux des interventions (y compris effets indirects, effets de second tour)	Une évaluation d'impact conçue comme une première étape d'une étude plus globale

...

...

Type d'utilisation	Utilisation précise	Que fait-on ?
Outil très opérationnel	Mesure de l'impact d'innovations du processus des interventions	Une étude mesurant l'effet causal d'une innovation dans la mise en œuvre d'un dispositif sur des variables intermédiaires (taux d'adhésion, etc.)
	Tests de modalités de mise en œuvre d'un projet	Une étude mesurant l'effet causal de différentes modalités de mise en œuvre du programme sur des variables intermédiaires (taux d'adhésion, etc.), afin de déterminer la modalité la plus efficace

Plusieurs utilisations possibles particulièrement utiles des évaluations d'impact, qui s'inscrivent dans un objectif de production de connaissances empiriques, sont évoquées par Parienté (2008) :

- il s'agit d'abord des évaluations d'impact de projets innovants (cf. GIZ, 2012) et/ou qui soulèvent des questions de recherche peu étudiées dans la littérature (*evaluation gaps*). Dans ce cas, les bénéfices espérés d'une évaluation d'impact paraissent particulièrement élevés, dans la mesure où le manque de connaissances (à la fois quantitatives et qualitatives) sur les effets attendus est avéré^[38].
- Les évaluations d'impact peuvent également être conçues dès l'origine pour analyser les facteurs de réussite/d'échec des projets (Deaton, 2010). Dans ce cas, ces évaluations seront d'emblée davantage centrées sur l'analyse de la chaîne de l'impact et de l'hétérogénéité (du contexte, de la mise en œuvre, des bénéficiaires), que sur la *mesure* des impacts, avec un protocole mixte privilégiant sans doute le volet qualitatif. On poursuivra là un objectif formatif ou d'apprentissage, qui n'était à l'origine pas forcément central dans les évaluations d'impact.
- Une autre utilisation utile possible consiste à concevoir les évaluations d'impact comme des tests empiriques d'hypothèses théoriques : par exemple, les travaux traitant de la question de l'élasticité de la demande de services ou de biens de base aux prix (J-Pal, 2011b), de l'importance du phénomène de sélection adverse sur le marché assurantiel (Levine *et al.*, 2012) ou encore de l'importance des externalités générées par un bien public (IRD – DIAL, 2013) paraissent parti-

[38] Exemple des études récentes sur le *mobile banking*.

culièrement intéressants (voir encadré 10). Cet usage va au-delà d'un simple intérêt académique, dans la mesure où ces hypothèses comportementales sous-tendent la plupart des programmes de développement. Par exemple, les travaux de Cadot *et al.*, (2013) mettent en évidence le fait que les interventions de l'AFD, sous la forme de lignes de crédit, de fonds de garantie ou de fonds d'investissement, reposent implicitement sur l'hypothèse de certaines défaillances de marché, qui n'ont dans les faits pas été testées empiriquement.

- Une autre utilisation prometteuse liée est l'usage des résultats de certaines évaluations d'impact microéconomiques dans le paramétrage de modèles d'équilibre général, afin de simuler les effets globaux des programmes. L'*International Food Policy Research Institute* (IFPRI) a ainsi réalisé un tel chantier dans le cadre de l'évaluation du programme *Progreso* au Mexique, afin de quantifier les effets indirects et de second tour de ce programme sur l'économie mexicaine (voir encadré 10). De façon plus générale, le couplage des évaluations d'impact avec des outils plus macroéconomiques paraît une piste encourageante pour tenter de mieux appréhender les effets de passage à l'échelle et de sortir, comme évoqué, du « mythe de l'effet tâche d'huile ».
- Enfin, l'utilisation des évaluations d'impact pour mener des analyses coûts – bénéfices des interventions dans le but d'identifier les interventions les plus efficaces pour atteindre un objectif donné est potentiellement intéressante, mais soulève néanmoins des questions sur la comparabilité des mesures des bénéfices utilisées et les limites déjà discutées d'utilisation directe trop naïve de tels classements.

Encadré 10 Quelques exemples de résultats novateurs des évaluations d'impact

Produits/services de base : faut-il faire payer l'utilisateur ?

Le J-Pal a publié en 2011 un policy Brief, basé sur les résultats d'une dizaine d'évaluations d'impact expérimentales qu'il a menées, et s'interrogeant sur la question controversée du bien-fondé de faire payer les usagers de produits ou de services de base (comme la santé ou l'éducation). Ces travaux concernaient les comprimés contre les vers intestinaux, les moustiquaires imprégnées, le chlore et les uniformes scolaires. Ces études aboutissent à des résultats qui paraissent particulièrement intéressants : ils établissent que la tarification, même faible, de ce type de produits en

...

...

réduit fortement l'achat (élasticité forte au prix) ; ils démontrent également que l'acte de payer n'augmenterait pas leur utilisation (mettant en cause l'idée intuitive du « *value for money* ») et que les personnes qui sont disposées à payer ne sont pas forcément celles ayant le plus à gagner (pas les plus malades par exemple ; pas de forte sélection adverse). Ces études montrent également que bénéficier de ce type de produit, dans un premier temps, gratuitement augmente la probabilité d'acheter le produit dans le futur (effet d'apprentissage sur les bénéfices du produit). Au total, pour les auteurs de ces travaux, la tarification des produits ne permettrait pas véritablement d'atteindre l'objectif de soutenabilité (mis en avant pour justifier la non-gratuité). Ils reconnaissent néanmoins des effets incitatifs favorables sur l'offre (sur les fournisseurs de ces produits). Il est clair que ces résultats, qui ont pu être établis grâce au caractère aléatoire de l'assignation du produit à différents prix, n'auraient pas pu être établis de façon crédible par d'autres méthodes d'évaluation.

Réconcilier analyse micro et analyse macro : l'exemple de l'utilisation de l'évaluation d'impact du programme *Oportunidades/Progres*a au Mexique

L'évaluation d'impact du programme *Oportunidades/Progres*a au Mexique est considérée comme la « *Rolls-Royce* » des évaluations d'impact dans le domaine du développement, en raison de ses résultats significatifs, qui ont influencé directement les décideurs politiques de ce pays comme des pays de la région (cf. *supra*). Plus récemment, l'IFPRI a publié les résultats d'une utilisation intéressante de ces travaux (Debowicz et Golan, 2013) : sur la base des données collectées dans le cadre de l'évaluation d'impact et des mesures microéconomiques d'impact faites, il a pu développer et utiliser un modèle de microsimulation combiné à un modèle d'équilibre général calculable, lui permettant de quantifier les effets macroéconomiques de ce programme. Les chercheurs de l'IFPRI établissent ainsi que les effets positifs d'*Oportunidades/Progres*a seraient encore plus importants d'un point de vue macroéconomique *i.e.* lorsque l'on prend en compte les effets indirects d'équilibre général. En particulier, l'attribution de prestations conditionnées à la fréquentation scolaire, en augmentant le coût d'opportunité du travail (ou en réduisant le coût d'opportunité d'étudier), réduirait l'offre de travail des enfants et donc augmenterait les salaires de ceux qui continuent de travailler (par raréfaction de l'offre). Par conséquent, ce programme augmenterait indirectement les revenus des ménages (pauvres) qui choisissent de maintenir leurs enfants au travail. Au total, en prenant en compte ces effets, l'impact du programme en termes de réduction de la pauvreté et des inégalités se trouverait accru (le taux de pauvreté est réduit de 1,8 point de pourcentage à l'équilibre partiel contre 2,7 points de pourcentage à l'équilibre général).

Les évaluations d'impact peuvent également être utilisées comme un outil très opérationnel, visant à optimiser la mise en œuvre des projets. L'évaluation d'impact permet ainsi de tester des modalités de mise en œuvre d'une intervention, afin de déterminer quelle est la plus efficace. C'est la stratégie adoptée par le programme DIME de la Banque mondiale (cf. encadré 11). Dans cette perspective, les mesures portent surtout sur des variables intermédiaires (adhésion/adoption, utilisation), plutôt que véritablement sur des variables de résultats ou d'impacts. Pour DIME, cette utilisation de l'évaluation d'impact en fait un outil prospectif permettant d'ajuster les interventions en « temps réel », afin d'espérer in fine maximiser les impacts produits (*impact evaluation for real time policy making*). Parienté (2008) évoque également, de façon liée, un usage des évaluations d'impact pour tester des innovations du processus d'interventions (déjà en place). Bien sûr, cette utilisation très opérationnelle des évaluations d'impact suppose une implication forte des opérateurs des projets/programmes.

Encadré 11 Focus sur la stratégie de DIME et exemples de résultats très opérationnels utiles

Stratégie de DIME

Le programme DIME de la Banque mondiale poursuit trois objectifs (Legovini, 2010) : améliorer la qualité des opérations de la Banque à travers un processus d'apprentissage itératif ; renforcer les capacités des pays partenaires en termes de politiques publiques basées sur la preuve et produire des connaissances globales sur des questions clés de développement.

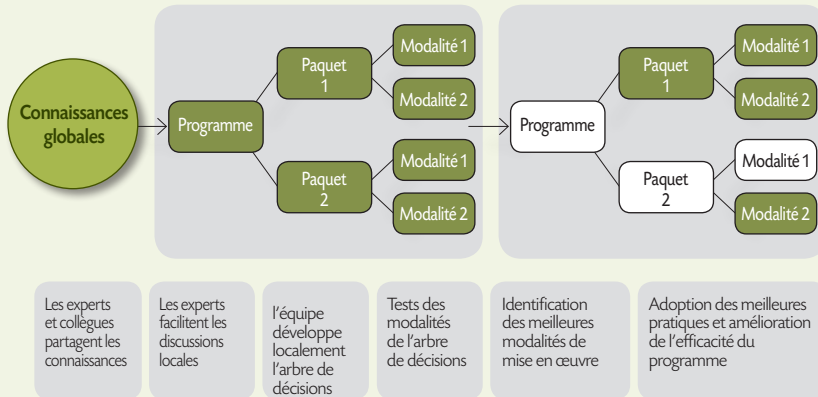
S'agissant du premier objectif, les équipes opérationnelles définissent avec l'appui des équipes de DIME des « arbres de décision » (*decision trees*), qui correspondent à des chaînes de résultats alternatives, à tester (cf. schéma 4). L'évaluation d'impact vise dans ce cadre à mesurer les effets de ces différentes modalités opérationnelles sur des variables intermédiaires de court terme (adhésion/adoption, utilisation). L'objectif poursuivi est donc de fournir des éléments de preuve très tôt dans le cycle du projet afin d'ajuster au mieux sa mise en œuvre et, ce faisant, tenter de contribuer à de meilleurs résultats à moyen terme. Ces évaluations d'impact prospectives visent à être le plus possible réalisées de façon conjointe avec les pays partenaires (à la fois concernant le choix des questions étudiées, la collecte et l'analyse des données). Ceci contribue au deuxième objectif de DIME de renforcer les capacités des pays partenaires (*learning by doing*), au-delà d'actions classiques de formation.

...

...

Schéma 4 *Modèle des évaluations d'impact de DIME : impact evaluations for real time policy making*

Améliorer l'efficacité des programmes



Source : Legovini, 2010.

Une illustration intéressante de ce type d'études est une expérimentation menée au Brésil (dans la ville de *Belo Horizonte*), visant à identifier la meilleure stratégie pour faire sortir les entreprises du secteur formel (De Andrade *et al.*, 2013). Ainsi, quatre modalités alternatives ont été testées : 1) une action d'information sur les procédures d'enregistrement ; 2) une action d'information couplée à la gratuité de l'enregistrement et l'accès à un comptable pendant un an ; 3) une action contraignante via la visite d'un inspecteur municipal et enfin 4) la visite d'un inspecteur municipal auprès d'une entreprise voisine, afin de capturer les éventuelles externalités d'une action contraignante. Les résultats établissent un impact nul voire négatif des actions incitatives (1 et 2), et un impact positif mais faible d'une action plus coercitive reposant sur la mise en place d'inspections. Ainsi, les auteurs concluent que, dans ce contexte particulier, les entreprises ne formalisent pas leur activité, sauf si elles sont forcées de le faire, ce qui suggère qu'elles retirent peu de bénéfices privés à l'enregistrement. Néanmoins, les revenus fiscaux supplémentaires tirés de cette formalisation accrue font plus que compenser le coût des inspections pour le gouvernement. Au-delà des résultats intéressants de cette étude, son point fort réside également dans le fait qu'elle ait été impulsée par le gouvernement de l'État de *Minas Gerais* dans lequel se situe *Belo Horizonte*, soucieux de ce que 72 % des entreprises de son territoire restaient informelles, malgré des décisions de simplification prises au niveau fédéral.

...

...

Évaluation d'impact d'un projet d'adduction en eau potable au Yémen

Il semble aller de soi qu'offrir un meilleur accès à l'eau potable à des populations va permettre d'améliorer significativement leur état de santé. Tel ne semble en réalité pas toujours le cas selon les résultats des évaluations d'impact sur ce thème (OCDE 2012b). En effet, d'autres facteurs (confondants) vont déterminer la pleine réalisation ou non de ces bénéfices sanitaires, comme notamment l'accès à l'assainissement et les pratiques d'hygiène. Une étude menée récemment au Yémen (Klasen *et al.*, 2012) remet en cause de façon encore plus troublante ce présupposé en mettant en évidence des impacts sanitaires *négatifs* d'une intervention visant à améliorer l'accès à l'eau potable (via la mise en place/l'extension de réseaux d'eau potable). Cet impact négatif s'explique par deux facteurs : l'irrégularité du service de distribution de l'eau, qui se traduit par une contamination de l'eau stagnante dans les réseaux et des pratiques d'hygiène des ménages restées inchangées (malgré l'augmentation de la quantité d'eau consommée). Les auteurs de l'étude en concluent que la mise en place (ou l'extension) de tels réseaux d'eau potable ne devrait être faite que lorsque la fiabilité du service de distribution est suffisamment bonne, sous peine de finalement générer des risques sanitaires supplémentaires. Cette étude, qui a allié un dispositif quasi-expérimental et un volet qualitatif permettant de remonter la chaîne des impacts, illustre tout l'intérêt opérationnel et académique de prévoir dès le départ un protocole intégré permettant d'identifier les facteurs de réussite/d'échec des projets.

3. Nouveaux enjeux posés par les évaluations d'impact

Comme on l'a vu dans la section 2, deux facettes des évaluations d'impact ont été largement débattues au cours des années récentes à la fois parmi la communauté des bailleurs et celle des chercheurs : il s'agit des *méthodes* des évaluations d'impact et de l'*utilisation* des résultats produits, thèmes sur lesquels, même si certains points de vue restent divergents, un certain degré de consensus semble avoir été atteint. En revanche, d'autres questions soulevées par ces travaux, dans le domaine du développement, semblent avoir reçu moins d'attention : il s'agit d'abord de la gouvernance de ces études dans le contexte de la Déclaration de Paris (cf. section 3-1), du mode de production des données qui en constituent le cœur (cf. section 3-2), de leur caractère supposé de « bien public » (cf. section 3-3), enfin, des perspectives de développement des évaluations d'impact dans un contexte international changeant (cf. section 3-4).

3.1. *Les évaluations d'impact sont-elles compatibles avec la Déclaration de Paris ?*

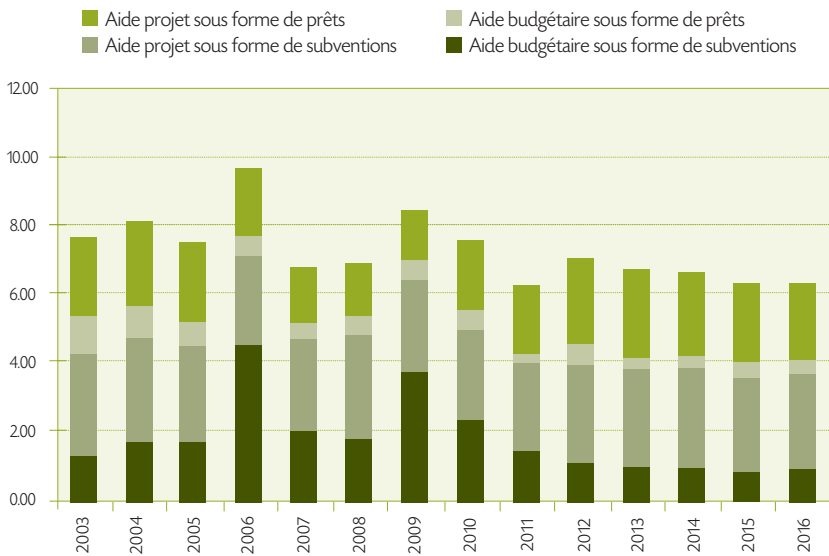
Les bailleurs se sont engagés, au travers successivement de la Déclaration de Paris (2005), des accords d'Accra (2008) et du partenariat de Busan (2011), sur certains principes d'efficacité de l'aide, tels que l'appropriation (par les pays bénéficiaires) et l'alignement (sur les institutions et les procédures des pays bénéficiaires). En particulier, selon le principe d'appropriation, les pays bénéficiaires de l'APD exercent une réelle maîtrise sur leurs politiques et stratégies de développement et assurent la coordination de l'action à l'appui du développement. Selon le principe d'alignement, les donateurs font reposer l'ensemble de leur soutien sur les stratégies nationales de développement, les institutions et les procédures des pays bénéficiaires de l'APD.

Or, certains travaux d'évaluation d'impact, qui exigent de modifier le mode opératoire « normal » des projets (par exemple, dans le cas des expérimentations), peuvent être perçus, dans une certaine mesure, en contradiction avec ces principes (sauf dans les cas – qui semblent minoritaires – où ce sont les décideurs politiques des pays partenaires qui sont à l'initiative de ces études). Ceci est particulièrement problématique dans les cas où la mise en œuvre des projets/programmes est encore fragile.

De façon liée, le *Congressional Research Service* (2013) craint que les évaluations d'impact reposant sur des protocoles très stricts viennent limiter la participation des bénéficiaires d'une intervention à sa conception (comme c'est le cas pour certains projets par exemple de développement rural).

Par ailleurs, l'on peut se demander si les évaluations d'impact accentuent la tension entre appropriation et évaluabilité. En ligne avec l'agenda sur l'efficacité de l'aide, des appels se sont fait jour pour accroître la part de l'aide publique délivrée sous forme d'aide budgétaire plutôt que d'aide projet. En effet, ce mode de délivrance est supposé avoir de multiples vertus : renforcement de l'appropriation de l'aide par les gouvernements des pays bénéficiaires (alors en charge de définir leurs priorités, de réaliser leurs arbitrages budgétaires et de gérer la mise en œuvre de leurs interventions) ; coordination accrue de l'aide des différents bailleurs (notamment au travers de fonds multibailleurs), réduisant la fragmentation de l'aide et les « coûts de coordination » associés. Ce faisant, ce mode d'intervention est potentiellement porteur de résultats de développement plus durables. Toutefois, le *Congressional Research Service* (*ibid.*) note que la promotion de cet outil réduit dans une certaine mesure la capacité des bailleurs à rendre compte avec précision de l'utilisation de leurs engagements et de leur contribution à un résultat de développement particulier. En témoignent les tentatives récentes de développement de méthodologies pour le moins complexes visant à évaluer les aides budgétaires (OCDE, 2013). Or, comme le constate les Nations unies (2013), l'aide budgétaire reste en réalité relativement limitée dans le portefeuille des bailleurs et plutôt en diminution au cours des deux dernières années (cf. graphique 9). Ce phénomène est précisément à relier, selon les Nations unies, aux pressions croissantes rencontrées par les bailleurs (notamment de leur Parlement) pour démontrer l'efficacité de l'aide octroyée, sur la base de résultats « mesurables » (plus faciles à établir pour des interventions ayant des objectifs bien ciblés *ex ante*). Même si les évaluations d'impact ne sont bien sûr pas responsables de ce phénomène, il n'en reste pas moins que, de par leur objet (mesurer les effets attribuables à une intervention) et leur champ d'application (des interventions de complexités faible à modérée), elles peuvent alimenter dans une certaine mesure la tendance naturelle des bailleurs à privilégier des modes d'intervention « simples » et ciblés de type aide projet.

Graphique 9 Répartition de l'APD par type d'instrument



Source : World Economic Situation and Prospects 2013, UN.

Enfin, la forte expansion des expérimentations au cours des dix dernières années peut être perçue comme autant de nouveaux « petits » projets de développement (dont une partie n'aurait peut-être pas été mise en place si ce n'était pour ces évaluations), risquant d'accentuer la fragmentation de l'aide, que justement la Déclaration de Paris visait à réduire. Pour illustration, l'IPA note, dans le cas du Ghana, qu'il est perçu par le gouvernement comme un interlocuteur additionnel parmi la communauté des bailleurs/ONG, ce qui est bien compréhensible compte tenu de la frontière floue évoquée entre activité de développement et activité d'évaluation de ce type d'opérateur. Autre illustration : l'évaluation d'impact de la micro-assurance de santé SKY au Cambodge, menée par l'AFD, est qualifiée par certains de « projet dans le projet » (Comité des évaluations de septembre 2012). Enfin, la production de données dans le cadre de ces études, le plus souvent via des enquêtes *ad hoc* externes aux systèmes statistiques nationaux, pose en soi des questions d'appropriation, d'alignement, et en termes de renforcement des capacités statistiques spécifiques (cf. engagements pris à Busan en 2011 ; cf. section suivante).

3.2. Comment mieux articuler production de données pour les évaluations d'impact et renforcement des systèmes statistiques nationaux ?

3.2.1. *Enquêtes ad hoc pour les évaluations d'impact versus enquêtes nationales : état des lieux*

L'expansion des évaluations d'impact dans le domaine du développement s'est accompagnée d'une expansion de la réalisation d'enquêtes *ad hoc* spécialement conçues et réalisées pour les besoins de ces études. Ainsi, dans le cas de la Banque mondiale, l'IEG (2012) note que dans environ 75 % des cas, les évaluations d'impact conduites ont nécessité des enquêtes spécifiques. À l'exception de quelques bailleurs (comme l'IDB), les évaluations d'impact menées au cours des dix dernières années semblent avoir souvent fait appel à une telle collecte de données dites primaires (même si ces données primaires sont souvent croisées avec des données secondaires/existantes ^[39]). De plus, compte tenu de la nature statistique des évaluations d'impact, ces enquêtes *ad hoc* de type quantitatif sont généralement réalisées sur un grand nombre de ménages (souvent plusieurs milliers ; cf. tableau 9), et ont donc de ce fait une envergure d'enquêtes nationales (même si elles ciblent généralement une zone géographique limitée).

[39] Ainsi, dans le cas de la Banque mondiale, seules environ 40 % des évaluations d'impact conduites ont reposé exclusivement sur des données primaires.

Tableau 9 Quelques exemples de taille d'échantillon dans le cadre d'enquêtes pour des évaluations d'impact et d'enquêtes nationales (pauvreté ou santé)

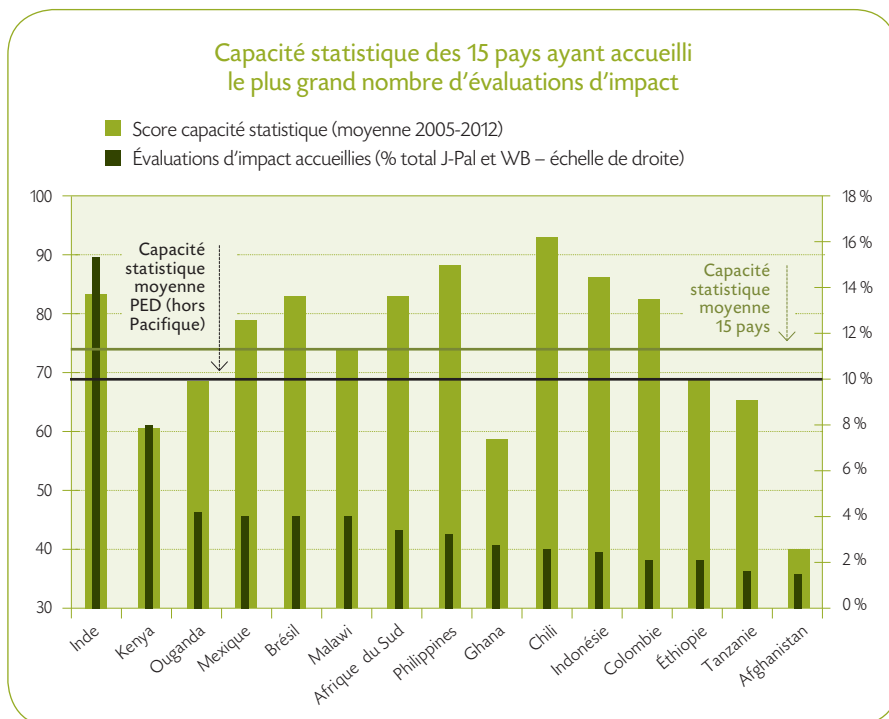
	Taille enquête étude d'impact	Taille enquête nationale (dernière menée)
Étude d'impact micro-assurance santé SKY au Cambodge (précitée)	5 213 ménages (2009)	15 667 ménages (Enquête démographie santé [EDS], 2010)
Étude d'impact microcrédit Al Amana Maroc (précitée)	5 551 ménages (2009)	7 200 ménages (enquête revenus et niveaux de vie, 2006-2007)
Étude d'impact santé Ouganda (Bjorkman et Svensson, 2009)	≈ 5 000 ménages (2006)	8 870 ménages (EDS, 2006)
Étude d'impact santé Kenya (Cohen <i>et al</i> , 2012)	2 789 ménages (2009)	8 561 ménages (EDS, 2003)*

* Une nouvelle vague a été réalisée en 2011 (rapport non disponible).

Or, le contexte dans lequel ces enquêtes sont réalisées est celui de pays dont le système statistique public est généralement fragile. Ainsi, si l'on considère les quinze pays ayant accueilli le plus grand nombre d'évaluations d'impact (selon les bases de données de la Banque mondiale et du J-Pal ; ces quinze pays représentant plus de 60 % des évaluations d'impact qui y sont recensées), ceux-ci ont certes une capacité statistique légèrement supérieure à celle de l'ensemble des pays en développement, mais l'on compte parmi eux des pays ayant un score très faible en termes de capacité statistique comme le Ghana (59 sur 2005-2012), le Kenya (61), sans parler de l'Afghanistan (40 ; cf. graphique 10). En outre, sur la période 2005-2012, deux de ces pays ont enregistré une nette diminution de leur capacité statistique (Brésil et Afrique du sud), tandis que deux, partant d'un niveau faible, ont connu une légère diminution ou une stagnation (Éthiopie et Kenya ; cf. tableau 10).

Graphique 10

Capacité statistique dans les pays ayant accueilli le plus grand nombre d'évaluations d'impact



Source : Banque mondiale, J-Pal.

Tableau 10

Évolution de la capacité statistique des 15 pays ayant accueilli le plus grand nombre d'évaluations d'impact entre 2005 et 2012 (score, 100 = maximum)

	2005	2012	Évolution
Inde	81	81	0
Kenya	58	58	0
Ouganda	69	70	1
Mexique	72	88	16
Brésil	86	76	-10
Malawi	66	74	9
Afrique du Sud	88	71	-17
Philippines	88	83	-4
Ghana	53	59	6
Chili	91	93	2
Indonésie	84	82	-2
Colombie	79	84	6
Éthiopie	64	63	-1
Tanzanie	63	69	6
Afghanistan	28	50	22
Moyenne 15 pays	71	73	2
Moyenne PED (hors Pacifique)	64	66	2

Source : Banque mondiale.

De façon plus générale, l'on note que si la capacité statistique globale des pays en développement a légèrement progressé entre 2005 et 2012, certaines fragilités importantes demeurent : en particulier, en Afrique subsaharienne, 17 % des pays n'avaient pas réalisé de recensements depuis plus de 10 ans en 2012, 72 % avaient des comptes nationaux dont l'année de base est très ancienne (supérieure à 10 ans), de même que pour l'indice des prix à la consommation (43 %)^[40], et la quasi-totalité de ces pays n'ont pas de système complet d'enregistrement des naissances et des décès (en 2009, la naissance de 50 millions d'enfants dans le monde n'a pas fait l'objet d'un enregistrement ; Busan, 2011). Si l'on s'intéresse plus précisément aux enquêtes pauvreté et santé (équivalentes – mais d'ampleur beaucoup plus large – des enquêtes ménages qui peuvent être réalisées pour les évaluations d'impact), l'on constate qu'environ un quart des pays en développement n'ont pas conduit de telles enquêtes depuis plus de 5 ans. En Afrique subsaharienne, si cette proportion est en recul s'agissant des enquêtes pauvreté (de 34 % en 2005 à 23 % en 2012), elle est en augmentation (quoiqu'à un niveau plus bas) pour les enquêtes santé (de 9 % à 13 %).

Si ces éléments ont été largement documentés (très récemment par Jerven, 2013), et que ces fragilités sont, dans certains cas, expliquées par des situations politiques conflictuelles, il n'en reste pas moins qu'ils posent question eu égard aux montants financiers significatifs investis au cours des dix dernières années dans le financement d'enquêtes spécifiques pour les évaluations d'impact.

[40] L'importance d'un rebasage suffisamment fréquent est notamment documentée par Jerven (2013), qui rappelle que le FMI préconise un rebasage tous les 5 ans (dans certains pays développés, les comptes nationaux sont passés depuis quelques années – comme en France – au système des prix chaînés). Ainsi, pour illustration, dans le cas du Ghana, suite au rebasage de ses comptes nationaux en 2010, le niveau estimé du PIB est passé de 21,7 Mds cedi à 36,9 Mds cedi !

Tableau 11 Quelques éléments illustrant la fragilité statistique persistante de nombreux pays en développement

	Nombre de pays en développement (entre parenthèses, en % du total)		
	2005	2012	Évolution (pts de %)
Pas de recensement depuis plus de 10 ans	23 (16)	17 (12)	-4
Dont Afrique subsaharienne	14 (30)	8 (17) Angola, RD Congo, Côte d'Ivoire, Érythrée, Guinée, Madagascar, Mauritanie, Somalie	-13
Année de base des comptes nationaux supérieure à 10 ans	81 (56)	81 (55)	0
Dont Afrique subsaharienne	32 (68)	34 (72)	+4
Année de base IPC supérieure à 10 ans	72 (50)	49 (34)	-16
Dont Afrique subsaharienne	27 (57)	20 (43)	-15
Pas d'enquête pauvreté depuis plus de 5 ans (LSMS, etc.)	39 (27)	36 (25)	-2
Dont Afrique subsaharienne	16 (34)	11 (23) Burundi, Comores, Djibouti, Guinée équatoriale, Érythrée, Gabon, Ghana, Lesotho, Somalie, Soudan, Yémen	-11
Pas d'enquête pauvreté depuis plus de 5 ans (LSMS, etc.)	39 (27)	40 (27)	+2
Dont Afrique subsaharienne	4 (9)	6 (13) Botswana, Djibouti, Guinée équatoriale, Érythrée, Gabon, Yémen	+4
Pas de système d'enregistrement des naissances/des décès complet	90 (62)	91 (62)	0
Dont Afrique subsaharienne	45 (96)	45 (96)	0
Capacité statistique globale (100 = maximum)	65	66	+1
Dont Afrique subsaharienne	56	58	+2

Source : Banque mondiale.

3.2.2. Financement des enquêtes pour les évaluations d'impact et aide à la statistique publique : y-a-t-il un effet de substitution ?

Si, comme on l'a vu, les évaluations d'impact et les enquêtes associées ont bénéficié de financements considérables depuis le milieu des années 2000, dans le même temps, l'aide consacrée par les bailleurs au financement des statistiques publiques semble également avoir nettement crû, notamment en lien avec l'adoption des OMD (passant d'environ 1 M USD en 2006-2008 à 2,3 Md USD en 2010-2012 selon Paris21 ; cf. graphique 11). Toutefois, il faut noter que cette aide est relativement concentrée puisque les 15 premiers pays bénéficiaires (sur 116) représentent environ la moitié de l'ensemble de ces engagements, et qu'elle n'est pas nécessairement orientée vers les pays ayant les capacités statistiques les plus faibles (ainsi, parmi ces 15 pays, seulement six ont un score de capacité statistique inférieur à la moyenne des pays en développement^[41]). En outre, deux pays, proches en termes de capacité statistique de la moyenne des pays en développement hors Pacifique^[42] (score de 70 en 2012), se distinguent en ayant accueilli un nombre élevé d'évaluations d'impact, mais une aide à la statistique publique relativement limitée (les financements d'enquêtes liées aux évaluations d'impact depuis 2005 pourraient être, selon nos estimations, supérieurs à l'aide à la statistique publique reçue sur 2010-2012) : il s'agit de l'Ouganda et de l'Afrique du sud. Le Kenya est également un exemple intéressant puisque, malgré le fait qu'il représente le onzième pays bénéficiaire de l'aide à la statistique publique, les financements alloués aux enquêtes menées dans le cadre des évaluations d'impact réalisées dans ce pays pourraient représenter de l'ordre d'un tiers de l'aide à la statistique publique reçue en 2010-2012 (ces chiffres doivent évidemment être pris avec beaucoup de précautions, cf. notes tableau 12). Ceci est d'autant plus saisissant que ce pays reste fragile en termes de capacité statistique (score de 58 en 2012).

Au total, ces éléments certes frustes (en lien avec les limites liées à la fois à la qualité des déclarations faites par les bailleurs auprès de Paris21 et à la difficile comptabilité des évaluations d'impact en termes de nombre et de budget) suggèrent qu'il n'y aurait pas d'effet de substitution global entre aide à la statistique publique et financements alloués aux enquêtes réalisées dans le cadre des évaluations d'impact. Néanmoins, dans le cas de certains pays, la présence d'un tel effet n'est pas à exclure. Or, si les enquêtes *ad hoc* financées pour mesurer l'impact d'un projet particulier ont un

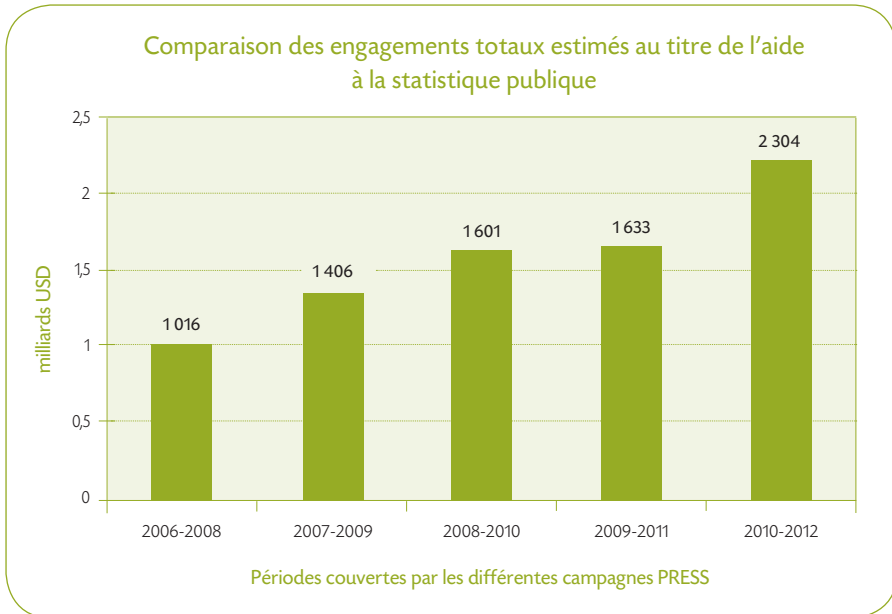
[41] Cette situation n'est cependant pas tout à fait surprenante puisque l'on retrouve pour l'aide à la statistique publique, la même tension – entre besoin et qualité de la gouvernance – que pour l'APD en général.

[42] La moyenne exclut cette zone, composée essentiellement de petites économies insulaires dont les capacités statistiques sont très faibles (score moyen de 47 contre 58 par exemple en Afrique subsaharienne).

bénéfice direct^[43], comme on l'a vu, souvent incertain ou diffus, les enquêtes menées dans le cadre de plans statistiques nationaux ont des bénéfices certains et très importants (*multi-purpose surveys*). Les données produites dans ce cadre fournissent des informations sur la situation nationale pour les administrations et les services publics qui sont utiles à la planification et au développement de programmes. Elles constituent également une source d'informations pour l'ensemble de la société civile (Rallu, 2010).

Graphique 11

Évolution de l'aide à la statistique publique au regard du développement des évaluations d'impact



Source : Banque mondiale, Paris 21, J-Paï, calculs auteur.

[43] Supposées être un « bien public », elles ont aussi vocation à générer des bénéfices indirects, point qui est toutefois discuté *infra*.

Tableau 12

Capacité statistique, aide à la statistique publique et évaluations d'impact au sein des 15 pays ayant accueilli le plus grand nombre d'évaluations d'impact

Pays	Score capacité statistique 2012	Aide à la statistique publique 2010-2012 (Ms USD)	Nombre d'évaluations d'impact depuis 2005 (limite basse *)	Budget estimé enquêtes **	Ratio (2)/(4)
	(1)	(2)	(3)	(4)	(5)
Inde	81	111,5	86	17,2	6,5
Kenya	58	43,4	40	8	5,4
Ouganda	70	4,4	25	5	0,9
Mexique	88	–	25	5	–
Brésil	76	–	25	5	–
Malawi	74	21	20	4	5,3
Afrique du Sud	71	2,4	21	4,2	0,6
Philippines	83	2,7	17	3,4	0,8
Ghana	59	48,9	17	3,4	14,4
Chili	93	–	16	3,2	–
Indonésie	82	67,9	8	1,6	42,4
Colombie	84	9,7	12	2,4	4,0
Éthiopie	63	36,8	13	2,6	14,2
Tanzanie	69	71,8	9	1,8	39,9
Afghanistan	50	63,6	9	1,8	35,3

Source : Banque mondiale, Paris 21, J-Pal, calculs auteur.

* Chiffre obtenu en cumulant les évaluations d'impact recensées sur le site de la Banque mondiale et du J-Pal, **tout en éliminant les doublons**. Il n'intègre pas des évaluations d'impact conduites dans ces mêmes pays par d'autres bailleurs et réalisées par d'autres centres de recherche.

** En faisant les hypothèses que les 2/3 des évaluations d'impact aient nécessité une collecte de données spécifique et que le coût unitaire moyen de cette collecte ait été de 300 000 USD (60 % x 500 000 USD ; IEG, 2012, Gertler *et al.*, 2011).

3.2.3. *Les instituts nationaux de statistique des pays partenaires sont-ils impliqués dans la collecte de données dans le cadre des évaluations d'impact ?*

Selon l'IEG (2012), environ 45 % des évaluations d'impact conduites par la Banque mondiale font appel à des chercheurs/enquêteurs locaux pour le travail de terrain et la collecte de données. Il ne fournit toutefois pas d'indications sur la nature précise de ces enquêteurs locaux. De façon plus générale, il ne semble pas que les bailleurs impliqués dans la conduite des évaluations d'impact, ni les principaux centres de chercheurs associés aient formalisé une stratégie volontariste en la matière. Pour exemple, l'IPA fait état de peu d'expériences d'externalisation de la conduite de ses enquêtes auprès d'instituts nationaux de statistique. Le J-Pal et l'IPA, très soucieux de maîtriser la qualité des données collectées, ont même une stratégie d'internalisation croissante des enquêtes réalisées (par du personnel propre à ces institutions). Tout au plus, la Banque mondiale (2007) émet-elle les recommandations suivantes quant aux étapes à franchir avant de décider de lancer une enquête spécifique dans le cadre d'une évaluation d'impact : vérifier s'il existe des données existantes pouvant être directement utilisées ; sinon, essayer de se « raccrocher » à une enquête déjà prévue, notamment par les instituts nationaux de statistique (en rajoutant un module de questions et/ou en accroissant la taille de l'échantillon) ; en dernier recours, lancer une enquête spécifique. Gertler *et al.* (2011) évoquent toutefois l'intérêt de commissionner la collecte de données aux instituts nationaux de statistique dans une optique de renforcement de capacités. Ils pointent néanmoins deux limites à cela : d'une part, les faibles capacités de ces instituts pour réaliser des activités additionnelles par rapport à leurs activités courantes ; d'autre part, leur potentiel manque d'expérience en termes d'enquêtes microéconomiques de panel. On peut néanmoins citer quelques exemples d'évaluations d'impact où la collecte des données a impliqué les instituts nationaux de statistiques (INS) : parmi les évaluations d'impact lancées par le MCC, au moins deux d'entre elles, se sont basées sur une collecte de données intégrée dans le cadre d'enquêtes nationales réalisées par les INS (cas d'études réalisées au Honduras et en Arménie) ; de même, au sein de l'AFD, deux évaluations d'impact se sont faites en collaboration avec les INS (à Djibouti et au Sénégal).

Comme effleuré dans la section précédente, l'externalisation, qui semble fréquente, de la production de données – en dehors des structures nationales publiques des pays partenaires – pose également la question de l'appropriation de ces données et de l'adéquation avec l'objectif de renforcement des capacités de ces structures, notamment réaffirmé par le partenariat de Busan en 2011. En effet, constatant à la fois les lacunes persistantes en termes de données (cf. *infra*) et la demande croissante

de redevabilité/de gestion axée sur les résultats, les partenaires de Busan appellent à un plan d'action en faveur du renforcement des capacités statistiques dans les pays en développement. L'OCDE (2012a) invite de son côté les pays donateurs, soucieux de disposer d'éléments tangibles sur les résultats générés par leurs financements, à d'une part s'appuyer sur les systèmes statistiques nationaux (et des enquêtes conduites par les pays dans le respect des normes internationales), d'autre part placer les évaluations (rigoureuses) sous l'égide des pays partenaires, avec l'appui d'institutions comme le 3IE.

3.2.4. *Quelles sont les possibilités effectives d'utilisation des données produites dans le cadre des évaluations d'impact pour le « public » ?*

Les évaluations d'impact ont vocation à être un « bien public » dont les résultats, mais aussi les outils et les bases de données sont partagés avec le public (que cela soit les décideurs politiques, les bailleurs, les statisticiens, les chercheurs ou la société civile). Outre que cette notion même de « bien public » peut être questionnée (cf. section suivante), l'on peut dans un premier temps s'interroger sur les possibilités *pratiques* d'utiliser les données produites dans le cadre de ces études. Par exemple, il est à craindre qu'un certain manque d'harmonisation des méthodes, des indicateurs et des nomenclatures utilisés (contrairement aux enquêtes/relevés statistiques faits dans le cadre de productions publiques nationales, qui sont soumis à des normes nationales / internationales) complique l'utilisation de ces données pour des usages autres que les évaluations pour lesquelles ils ont été produits, et surtout les comparaisons avec d'autres sources (cf. encadré 12). En outre, il est également à craindre que ces données produites de façon « privée » soient insuffisamment documentées pour que l'on puisse aisément juger de leur qualité (manque de métadonnées et de guides méthodologiques d'accompagnement). Cet élément doit toutefois être modéré dans le cas de certaines évaluations d'impact de la Banque mondiale : en effet, DIME essaie d'harmoniser le plus possible les questionnaires et les indicateurs utilisés afin de favoriser les comparaisons entre pays notamment (Legovini, 2010).

Encadré 12 Comparabilité d'une enquête ad hoc par rapport à une enquête nationale : l'exemple d'une évaluation d'impact au Maroc

La comparaison du module « dépenses et revenus du ménage », utilisé dans le cadre d'une évaluation d'impact au Maroc, et du questionnaire standard établi par la Direction de la statistique du Maroc pour l'enquête nationale sur la consommation et les dépenses des ménages de 2006-2007 permet de mettre en évidence des différences non négligeables :

- l'ensemble des dépenses de consommation ne sont pas couvertes dans le module de l'enquête pour l'évaluation d'impact, par rapport à celui utilisé dans les enquêtes ménages officielles au Maroc ;
- certaines dépenses énumérées dans le questionnaire de l'étude d'impact ne sont pas comptablement des dépenses de consommation : dons à l'Imam ; remboursement de crédit ;
- certains postes, non directement rattachés à la nomenclature des biens et services, ont un contenu potentiellement mal défini (exemple : loisirs) ;
- la période de rappel/mémoire ne suit pas les mêmes normes que celles utilisées par la Direction de la statistique marocaine ;
- l'autoconsommation inclut principalement les dépenses de consommation alimentaire (alors qu'elle inclut l'ensemble des dépenses de consommation dans le questionnaire de l'enquête nationale du Maroc).

...

Questionnaire enquête évaluation d'impact AI Amana (module dépenses des ménages)

SECTION H : DEPENSES ET REVENUS DU MENAGE

Montant En Dhs	H3. Combien avez-vous <u>dépensé</u> le MOIS dernier en :	H2. Combien avez-vous consommé la SEMAINE dernière des produits suivants de votre production ?	H1. Combien avez-vous dépensé la SEMAINE dernière pour les produits suivants ?	INSTRUCTION ENQUÊTER : pour répondre à Question H2, sur l'autoconsommation, utiliser feuille séparée : « Calcul de l'Autoconsommation ».
		Montant En Dhs		
	Eau	1		Farine
	Electricité	2		Légumes
	Téléphone (fixe et cellulaire)	3		Fruits et noix
	Bougies et autres produits pour l'éclairage	4		Lait ou produits laitiers
	Gaz (Butane) et autres combustibles	5		Pain
	Savon et autres pour utilisation personnelle	6		Huile
	Produits de propreté (savon et autres détergents) pour la maison	7		Thé
	Frais médicaments et santé	8		Café
	Loyer de la maison	9		Sucre
	Remboursement crédit	10		Poulet
	Dons à l'Imam	11		Viande
	Journaux, magazines	12		Œufs
	Location de biens de la maison	13		Poisson
	Loisirs	14		Epices
	Lister autres dépenses de consommation du mois dernier			Transport
		15		Cigarettes
		16		Repas pris à l'extérieur
		17		Boissons prises à l'extérieur
	Total dépense par mois	18		Lister autre dépenses/autoconsommation de la semaine dernière
				19
				20
				21
				Total dépense par semaine
				22

H5. Financement 1=Epargne ; 2=Crédit ; 3=Dons ; 4=Autre ; 97/99=NR/NS	H4. Montant En Dhs	Quelles ont été vos dépenses l'ANNEE dernière en :
		Frais scolaires (inscription, rentrée et dépenses pendant l'année)
		Habits
		Chaussures
		Construction ou réparation de la maison
		Frais du mois de ramadan
		Fête du mouton (Aïd el Kabir)
		Voyages (en dehors de la commune)
		Dons et transferts à d'autres ménages (famille/amis)
		Lister autres dépenses de l'année dernière
		9
		10

...

H7. Financement 1=Epargne ; 2=Crédit 3=Don ; 4=Autre. 97/98=NE/NS	H6. Montant En Dhs	Quelles ont été vos dépenses occasionnelles ou imprévues l'année dernière en :	
		Pèlerinages	1
		Naissance	2
		Circoncision	3
		Mariage/ fiançailles d'un fils ou d'une fille	4
		Frais de santé à la suite de la maladie d'un membre du ménage	5
		Décès d'un membre du ménage	6
		Domage de la maison (feu, inondation, autres)	7
		<i>Lister autres dépenses occasionnelles ou imprévues</i>	
			8
			9
Dhs	Quelles sont les dépenses habituelles, hebdomadaires, en nourriture ?		H8
Dhs	Combien dépensez vous au total, en nourriture et autres dépenses du ménage, par mois ? Inclure seulement dépenses qui se réalisent tous les mois (items tableau H1 et H3)		H9

Au total, il paraît donc indispensable de définir une stratégie explicite de meilleure articulation entre production de données dans le cadre des évaluations d'impact (le cas échéant) et soutien aux systèmes nationaux de statistique. En outre, de façon incidente, l'IEG World Bank (2009) note que l'implication des instituts nationaux de statistiques augmente la crédibilité et donc l'appropriation des résultats des évaluations d'impact par les gouvernements des pays partenaires.

3.3. Les évaluations d'impact : un bien public ou un bien de club ?

Deux arguments sont fréquemment évoqués pour justifier le coût souvent élevé des évaluations d'impact : l'utilisation de ces études pour arrêter/modifier des programmes inefficaces ou au contraire répliquer/étendre des programmes efficaces, dont on a vu qu'elle n'était pas toujours avérée ; la dimension « bien public » de ces études. C'est cette dimension de « bien public » qui est ici discutée. Qu'est-ce qu'un « bien public » ? Tel que défini en sciences économiques, il s'agit d'un bien ou d'un service dont l'utilisation est non-rivale et non-exclusive. La non-rivalité implique que la consommation du bien par un agent n'a aucun effet sur la quantité disponible de ce bien pour les autres individus. La non-exclusion implique qu'une fois que le bien public est produit, tout le monde peut en bénéficier. Il existe par croisement de ces deux critères trois autres catégories de biens : le bien de club (non-rival exclusif), le bien commun ou bien public impur (rival non-exclusif) et le bien privé (rival exclusif).

Afin de caractériser les évaluations d'impact par rapport à cette notion de bien public, 3 éléments constitutifs de celles-ci seront considérés : il s'agit d'abord des outils (questionnaires, méthodes, etc.) créés pour ces études, puis des données collectées^[44], enfin, des résultats et des éventuelles recommandations qui sont établis à partir de ces travaux. Ainsi définies, les évaluations d'impact semblent effectivement respecter le critère de non-rivalité : dès lors que les outils, les données et les résultats constitutifs de ces travaux sont « consommables », le fait qu'un agent les mobilise n'en réduit pas la disponibilité pour les autres. En revanche, le critère de non-exclusion semble davantage poser problème : en effet, la mise à disposition effective des outils créés pour ces études et des données collectées semble de fait limitée. Par exemple, sur 420 évaluations d'impact recensées dans la base du J-Pal, seulement 21 possèdent une base de données (et des outils associés comme les questionnaires) mise en ligne, soit environ 5 % du total. IEG (2012) note de son côté que seules 5 % des données collectées sur la période 2000-2009 par la Banque mondiale, soit dans le cadre d'évaluations d'impact, soit dans le cadre du suivi-évaluation des projets sont accessibles, alors que dans le même temps la Banque mondiale a consacré 419 M USD au financement de ces données. Le personnel même de la Banque déplore cet accès limité (même en interne). Dans le seul cas des évaluations d'impact, on peut estimer à 9 % la part des études conduites par la Banque mondiale pour lesquelles les bases de données (et les outils associés) sont accessibles au travers du *Data catalogue* de

[44] Que ces données soient des données primaires ou secondaires.

la Banque. Parmi les bailleurs examinés, seul le MCC se démarque en publiant sur son site les outils et les bases de données de façon plus systématique. De façon liée, on peut noter que ni les principaux bailleurs actifs en matière d'évaluations d'impact, ni les centres de recherche associés ne semblent avoir de politique explicite en matière de *délai* de diffusion de ces données. Enfin, s'agissant des résultats et recommandations issus de ces travaux, même si un effort important de vulgarisation est fait, il n'en reste pas moins que leur pleine compréhension nécessite la présence d'une fonction de *knowledge broker* au sein des institutions vers lesquelles ces résultats/recommandations sont destinés (notion de péage). Au total, il semblerait donc que jusqu'à présent, les évaluations d'impact s'apparentent plutôt à un bien de club. Signalons en outre que si les chercheurs des pays partenaires sont souvent associés à ces études au travers de la collecte des données (dans 45 % des études de la Banque mondiale par exemple ; IEG, 2012), ceux-ci ne sont généralement pas les principaux contributeurs/auteurs des analyses et études publiées (dans moins de 20 % des cas s'agissant de la Banque mondiale, selon la même source).

Tableau 13 *Les évaluations d'impact sont-elles un bien public ?*

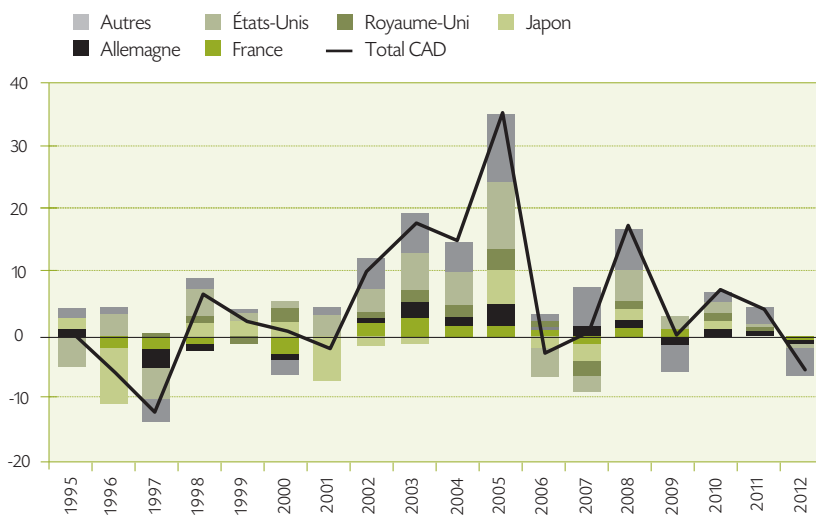
	Non-rivalité	Non-exclusion
Outils : questionnaires, méthodes, etc.	Oui	Accessibilité effective ? Délai important/asymétrie d'information (pas d'accès toujours aisé aux outils/données même lorsqu'ils sont diffusés)
Données collectées	Oui	
Résultats et recommandations	Oui	Péage : disposer de fonctions de <i>knowledge brokers</i> au sein des institutions pour avoir le plein bénéfice de ces résultats

3.4. Quelles perspectives de développement des évaluations d'impact dans un contexte international changeant ?

L'essor marqué des évaluations d'impact dans le développement est intervenu dans le contexte économique relativement favorable du milieu des années 2000. Toutefois, depuis 2008, les pays de l'OCDE ont été confrontés à la crise financière et économique la plus sévère enregistrée depuis la fin de la seconde guerre mondiale. En conséquence, la plupart de ces pays sont actuellement engagés dans des politiques d'ajustement budgétaire très contraignantes. D'ailleurs, en 2012, l'APD (telle que comptabilisée par le CAD) a diminué (en % du RNB) pour la deuxième année consécutive (ce qui n'était pas arrivé depuis le milieu des années 1990). En outre, la totalité des principaux contributeurs à l'APD a participé au recul observé en 2012 (ce qui n'était pas arrivé par le passé, même en 1997). Ce contexte dégradé peut donc conduire à s'interroger sur la durabilité du soutien budgétaire conséquent dont ont bénéficié les évaluations d'impact jusqu'à présent, d'autant plus qu'en 2012, la baisse de l'APD s'est faite en grande partie au détriment des institutions multilatérales (souvent les plus actives en la matière, comme on l'a vu). D'ailleurs, même si certains pays ont réaffirmé récemment leur soutien financier aux évaluations d'impact (États-Unis via le DIV, Royaume-Uni via le SIEF et le 3IE), il semble y avoir des incertitudes sur le financement du 3IE après 2013 (ODI et Tripeline consulting, 2012).

Graphique 12 Une réduction historique de l'APD en 2012

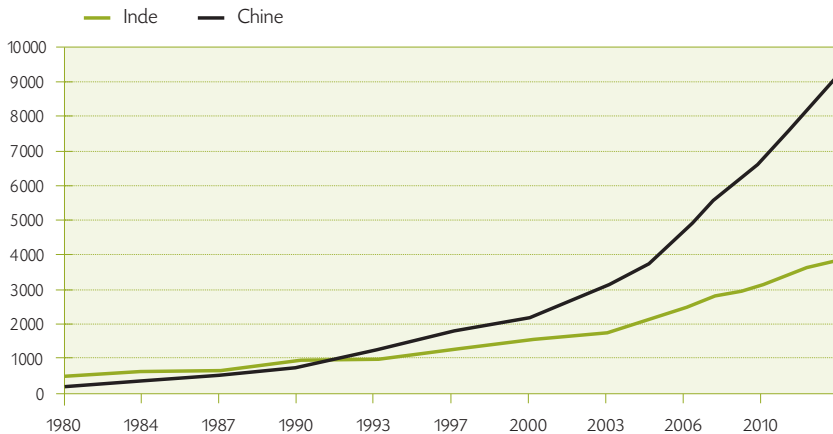
Contributions des principaux pays donateurs à la croissance de l'APD
(Points de % du total de croissance de l'APD totale des pays du CAD)



Source : CAD, calculs auteur.

Par ailleurs, cette crise financière et économique, de même que l'émergence depuis plusieurs années de nouveaux pays dans le concert des nations, rappellent l'importance de ne pas négliger les grandes questions macroéconomiques, auxquelles les évaluations d'impact ne permettent pas de répondre (ressorts du décollage ou non de certains pays : Chine, Inde, etc., cf. graphique 13 ; risques macroéconomiques encourus suite aux conséquences de la crise financière, etc.). De fait, en dépit de la promotion très forte qui est faite de certains résultats des évaluations d'impact, Rosenzweig (2012) note que les impacts, même positifs qui y sont mesurés, sont souvent marginaux, et insuffisants pour espérer enclencher une dynamique de sortie de la pauvreté des populations étudiées (« [...] *none of the effects discussed in Poor Economics*^[45] make a perceptible dent in reducing global inequalities »). Ravallion (2012) note pour sa part que les débats sur l'efficacité de l'aide ou la demande de redevabilité portent plutôt sur le portefeuille global des financements des bailleurs que sur des projets individuels. Ruben (2012) de son côté évoque les questions clés de complémentarité/substituabilité des interventions, qui appellent également des approches plus globales. Picciotto (2012) fait d'ailleurs le pari que la « bulle » actuelle en faveur de l'expérimentation dans le domaine du développement va finir par éclater (« *The current bubble of enthusiasm in development evaluation is bound to burst* »). Au total, ces auteurs nous invitent à considérer l'agenda de la recherche de solutions microéconomiques au phénomène de pauvreté comme un complément, et non un substitut, à une quête de solutions plus globales.

Graphique 13 *Le décollage économique de la Chine illustre l'importance de ne pas négliger les outils d'analyse macroéconomique (PIB par tête, PPA USD)*



[45] L'ouvrage « grand public » de référence sur les expérimentations écrit par Banerjee et Duflo (2009).

Annexes

1. Sources utilisées

Il n'existe pas de base unifiée recensant l'ensemble des évaluations d'impact réalisées (ou en cours) dans le domaine du développement, même si un projet en ce sens conduit par le 3IE est actuellement en cours (<http://www.3ieimpact.org/blog/2013/05/06/impact-evaluation-databases-can-you-find-everything-youre-looking/>).

Dans le papier, les bases suivantes ont été privilégiées :

- 3IE (761 évaluations d'impact recensées dont environ 130 financées par le 3IE) ;
- Banque mondiale (483 évaluations d'impact) ;
- J-Pal (397 évaluations d'impact) ;
- CEGA (près de 100 évaluations d'impact).

Il s'agit à la fois de bases des principaux bailleurs à l'origine de l'augmentation des évaluations d'impact dans le domaine du développement (Banque mondiale, 3IE) et de centres de recherche leaders dans ce domaine (J-Pal ; CEGA). A noter que les mêmes études peuvent être comptabilisées dans plusieurs de ces bases compte tenu des financements et/ou collaborations croisées.

Il est surprenant de constater que la plupart de ces bases présentent des limites importantes :

- à l'exception de celle de la Banque mondiale, elles ne sont pas disponibles (même sur demande) en version Excel, ce qui rend leur analyse complexe ;
- la base de la Banque mondiale n'est pas à jour après 2010 ;
- la base du J-Pal semble présenter certaines incohérences dans la comptabilisation du nombre d'évaluations réalisées ou en cours.

2. Liste des institutions interviewées

Les institutions ayant pu être interviewées dans le cadre de cette étude sont les suivantes :

- KfW
- GIZ
- DFID
- JICA
- USAID
- BEI
- Commission européenne
- IDB
- Nations unies (UNEG)
- J-Pal
- IPA
- 3IE

3. Quelques prérequis relatifs à la faisabilité et à l'efficacité d'une expérimentation aléatoire contrôlée

La faisabilité *a priori* de l'utilisation de la méthode RCT exige qu'un certain nombre de prérequis soient satisfaits, qui sont pour certains énoncés par Melia (2014) :

Prérequis techniques :

- par définition, l'expérimentation doit être conçue à minima avant la mise en œuvre du projet et le choix des unités bénéficiaires (zones, institutions, individus, etc.), au mieux, de concert avec la conception du projet ; en pratique, pour un bailleur, cela exige donc que l'expérimentation soit définie tout au démarrage de l'instruction d'un projet ;
- la population statistique éligible (bénéficiaires potentiels, contrôles potentiels) au projet (individus, organisations, etc.) doit être de taille suffisante ; ainsi, cette méthode assure des mesures sans biais des impacts sous couvert d'échantillons de taille assez grande (généralement plusieurs milliers d'individus^[46]). En particulier, elle ne peut pas être utilisée si l'intervention cible une population très particulière qui risque d'être presque entièrement servie par cette intervention (plus de contrôle de taille suffisante valide) ; *a contrario*, en cas d'adhésion volontaire à un programme, celui-ci doit être suffisamment attractif faute de disposer d'un groupe de bénéficiaires assez important ;
- l'intervention ne doit concerner qu'une fraction réduite de la population totale (sauf à exploiter l'éventuel phasage dans le temps de la mise en œuvre) ; en d'autres termes, l'échelle du projet doit être plutôt locale que globale/nationale.

Prérequis éthiques/politiques :

- l'expérimentation ne doit pas conduire à exclure certains individus qui auraient bénéficié de l'intervention (qui est supposée *a priori* bénéfique) ;
- le principe du tirage aléatoire doit être accepté par les différentes parties prenantes au projet.

[46] Même si la taille minimale des échantillons varie en fonction de la taille des impacts recherchée et de la nature des variables étudiées.

Prérequis pratiques :

- des programmes/des chocs externes au projet évalué ne doivent pas affecter un groupe (bénéficiaires ou contrôle) et non un autre (contrôle ou non-bénéficiaires) ; en pratique toutefois, cette contrainte peut être contournée en utilisant des variables de contrôle (même si cela se traduit alors par des mesures moins fiables des impacts).

D'autres facteurs de réussite sont également à prendre en compte dans la décision de lancer une expérimentation aléatoire contrôlée – même si ces facteurs ne sont malheureusement pas toujours connus *a priori* (voir notamment Bernard *et al*, 2012) :

- compte tenu des difficultés de maintenir l'intégrité d'une expérimentation sur une longue période, les effets attendus de l'intervention doivent se matérialiser assez rapidement, ce qui implique soit de s'intéresser aux interventions dont les chaînes de causalité sont courtes, soit de se concentrer sur les premiers résultats des interventions (plutôt que sur leurs impacts au sens du CAD). Ceci implique également que l'intervention testée soit suffisamment « mûre » ;
- de façon liée, le temps de la mise en œuvre du projet ne doit pas être trop long, faute de quoi l'échantillon de référence sélectionné pourra se trouver compromis (difficile à retrouver) ;
- l'intervention évaluée doit être relativement simple ; en cas d'interventions complexes, l'évaluateur risque d'être confronté (outre des problèmes de faisabilité) à des difficultés d'interprétation des résultats.

Liste des sigles et abréviations

ADB	<i>Asian Development Bank</i>
AFD	Agence Française de Développement
AfDB	Banque africaine de développement
APD	Aide publique au développement
AusAid	<i>Australian Agency for International Development</i>
BEI	Banque européenne d'investissement
BOP	<i>Bottom of the pyramid</i>
CAD	Comité d'aide au développement
CCT	<i>Conditional Cash Transfer</i>
CDI	<i>Centre of Development Impact</i>
CE	Commission européenne
CEGA	<i>Center for Effective Global Action</i>
CGD	<i>Center for Global Development</i>
CLEAR	<i>Center for Learning, Evaluation and Results</i>
DFID	<i>Department for International Development</i>
DIME	<i>Development Impact Evaluation Initiative</i>
DIV	<i>Development Innovation Ventures</i>
EDS	Enquête démographie santé
EES	<i>European Evaluation Society</i> (Société européenne d'évaluation)
F3E	Fonds pour la promotion des études préalables, des études transversales et des évaluations
GDIV	<i>Global Development Innovation Ventures</i>
GIZ	<i>Gesellschaft für Internationale Zusammenarbeit</i>

JEL	<i>Journal of Economic Literature</i>
IBRD	<i>International Bank for Reconstruction and Development</i>
IDA	<i>International Development Association</i>
IDB	<i>Inter-American Development bank</i>
IDS	<i>Institute of Development Studies</i>
IEG	<i>Independent Evaluation Group</i>
IETF	<i>Impact Evaluation Task Force</i>
IFPRI	<i>International Food Policy Research Institute</i>
INS	Institut national de statistiques
IOB	<i>Inspectie Ontwikkelingssamenwerking en Beleidsevaluatie</i>
IPA	<i>Innovations for Poverty Action</i>
IPC	Indice des prix à la consommation
ISC	<i>International Child Support</i>
JEL	<i>Journal of Economie Literature</i>
J-Pal	<i>Abdul Latif Jameel Poverty Action Lab</i>
JICA	<i>Japan International Cooperation Agency</i>
LSMS	<i>Living Standards Measurement Study</i>
MAP	Modernisation de l'action publique
MCC	<i>Millenium Challenge Corporation</i>
MED	<i>Monitoring and Evaluation Directorate</i>
MICS	<i>Multiple Indicator Cluster Survey</i>
NONIE	<i>Network of Networks for Impact Evaluations</i>
OCDE	Organisation de coopération et de développement économiques
ODI	<i>Overseas Development Institute</i>
OMD	Objectifs du Millénaire pour le développement

ONG	Organisation non gouvernementale
PEGNet	<i>Poverty Reduction, Equity and Growth Network</i>
RCT	<i>Randomized Controlled Trial</i>
RSA	Revenu de solidarité active
SFE	Société française d'évaluation
SIEF	<i>Strategic Impact Evaluations Fund</i>
UE	Union européenne
UNEG	<i>United Nations Evaluation Group</i>
USAID	<i>United States Aid for International Development</i>
USD	Dollar américain
3IE	<i>International Initiative for Impact Evaluations</i>

Bibliographie

AUSAID (2011), "3IE and the Funding of Impact Evaluations – a Discussion Paper for 3IE's Members", juillet.

BANERJEE, A. et E. DUFLO (2009), "The Experimental Approach to Development Economics", *Annual Review of Economics, Annual Reviews*, vol. 1 (1), pp. 151-178, Mai.

BANQUE ASIATIQUE DE DÉVELOPPEMENT (2010), "Implementing Impact Evaluation at ADB", *Technical Assistance Report*, décembre.

BANQUE MONDIALE (2009), "Making Smart Policy: Using Impact Evaluation for Policy Making – Case Studies on Evaluations that Influenced Policy", *WB Doing impact Evaluation*, No.14, Juin.

BANQUE MONDIALE (2007B), "Data for Impact Evaluation", *Doing Impact Evaluation*, No.06, octobre.

BERNARD, T., J. DELARUE et J.-D. NAUDET (2012), "Impact Evaluations: a tool for Accountability? Lessons from Experience at Agence Française de Développement", *Journal of Development Effectiveness*, volume 4, issue 2, pp. 314-327.

BILLÉ, R. (2009), « Agir mais ne rien changer ? De l'utilisation des expériences pilotes en gestion de l'environnement », *VertigO – la revue électronique en sciences de l'environnement*, Débats et Perspectives, septembre.

BJORKMAN, M. et J. SVENSSON (2009), "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda", *Quarterly Journal of Economics*, No. 124 (2).

BRILLION, N. (2013), « Bilan des évaluations décentralisées réalisées par l'AFD en 2010 et 2011 », *Ex post Evaluation et capitalisation*, AFD, n° 49, septembre.

BUSAN (2011), "Statistics for Transparency, Accountability and Results: a Busan Action Plan for Statistics", novembre.

CADOT, O., A-C. DISDIER, J. GOURDON, J. HÉRICOURT et A. SUWA-EISENMANN (2013), « Évaluer l'impact des instruments financiers en faveur des entreprises », *Rapport interne*, juin.

CAE (2013), « Évaluation des politiques publiques », *Les notes du Conseil d'analyse économique*, n° 1, février.

CAFFERINI, L. (2011), « Bilan des évaluations de projet réalisées par l'AFD entre 2007 et 2009 », *Ex post Evaluation et capitalisation*, AFD, n° 45, décembre.

CAIRNCROSS, S. (2012), *Présentations internes AFD*, avril.

CAIRNCROSS, S. et V. VALDMANIS (2006), "Water Supply, Sanitation and Hygiene Promotion", dans "Disease Control Priorities in Developing Countries", chapitre 41.

CAMFIELD, L. (2013), "Qualitative Methods in Modern Impact Evaluation – Ensuring Rigour and Reliability", Presentation at the CDI Seminar "Impact, Innovation and Learning: Towards a Research and Practice Agenda for the Future", 26-27 mars.

CAMPOS, F., A. COVILLE, A.M. FERNANDES, M. GOLDSTEIN et D. MCKENZIE (2012), "Learning from the Experiments that Never Happened – Lessons from Trying to Conduct Randomized Evaluations of Matching Grant Programs in Africa", *World Bank Policy Research Working Paper*, n° 6296, décembre.

CGD (2006), "When will we ever Learn? Improving Lives through Impact Evaluation", *Report of the Evaluation Gap Working Group*, mai.

CLEAR (2013), « Étude de la demande de renforcement des capacités en évaluation pour l'Afrique francophone », Présentation au séminaire de la SFE « L'évaluation du développement : quelles perspectives ? », février.

CLEAR (2012), "African Monitoring and Evaluation Systems – Exploratory Case Studies: a Collection of Case Studies facilitated by the Clear Initiative – WITS", septembre.

COHEN, J., P. DUPAS, et S. SCHANER (2012), "Price Subsidies, Diagnostic Tests and Targeting of Malaria Treatment: Evidence from a Randomized Controlled Trial", *NBER Working Paper*, No. 17943, mars.

CONGRESSIONAL RESEARCH SERVICE (2013), "Does Foreign Aid Work? Efforts to Evaluate US Foreign Assistance", *CRS Report for Congress*, février.

CRÉPON, B., E. DUFLO, M. GURGAND, R. RATHELOT et P. ZAMORA (2011), « L'accompagnement des jeunes diplômés demandeurs d'emploi par des opérateurs privés de placement : les enseignements d'une évaluation », *DARES Analyses*, n° 94, décembre.

CROZET, M., T. MAYER, F. MAYNERIS et D. MIRZA (2013), « Efficacité des dispositifs de soutien à l'exportation des firmes françaises », *Rapport d'étude du CEPII*, n° 03, mars.

DE ANDRADE, G., M. BRUHN et D. MCKENZIE (2013), "A Helping Hand or the Long Arm of the Law? Experimental Evidence on what Governments can do to Formalize Firms", *World bank Policy Research Working Paper*, No.6435, mai.

DEATON, A. (2010), "Instruments, Randomization and Learning about Development", *Journal of Economic Literature*, 48, pp. 424-455, Juin.

DEBOWICZ, D. et J. GOLAN (2013), "The Impact of Oportunidades on Human Capital and Income Distribution", *IFPRI Discussion Paper*, No.01257, mars.

DERCON, S. (2013), "Acting on Evidence – the Case of Microfinance", Presentation at CGAP Seminar on "Impact Evaluation for Financial Inclusion: A Retreat for CGAP Members with Leading Researchers", 10-11 janvier.

DFID (2013), "International development evaluation policy", mai.

DIAMOND, A. (2013), "An Introduction to Impact Evaluation in Observational (non-experimental) Settings", Presentation at CGAP seminar on "Impact Evaluation for Financial Inclusion: A Retreat for CGAP Members with Leading Researchers", 10-11 janvier.

DUFLO, E. et M. KREMER (2003), "Use of Randomization in the Evaluation of Development Effectiveness", Paper prepared for the World Bank Operations Evaluation Department (OED) Conference on Evaluation and Development Effectiveness, janvier.

DUFLO, E., R. GLENNERSTER et M. KREMER (2006), "Using Randomization in Development Economics Research: a Toolkit", *NBER Technical Working Paper*, No.333, décembre.

DURAND, C. et C. NORDMANN (2011), « Misère de l'économie du développement », *La Revue des livres*, no. 1, octobre.

DZAKPASU, S., T. POWELL-JACKSON et O. CAMPBELL (2013), "Impact of User Fees on Maternal Health Service Utilization and Related Health Outcomes: a Systematic Review", *Health Policy and Planning*, pp. 1-14, janvier.

FINKELSTEIN, A. (2005), "The Aggregate Effects of Health Insurance: Evidence Form the Introduction of Medicare", *NBER Working Paper Series*, No.11619, septembre.

FROT, E. et J. SANTISO (2008), "Development Aid and Portfolio Funds: Trends, Volatility and Fragmentation", *OECD Working paper*, No.278, décembre.

GAARDER, M. et B. BRICENO (2010), "Institutionalisation of Government Evaluation: Balancing Trade-Offs", *3IE Working Paper*, No.8, juillet.

GERTLER, P., S. MARTINEZ, P. PREMAND, L. RAWLINGS et C. VERMEERSCH (2011), "Impact Evaluation in Practice", Interactive Textbook at <https://worldbank.org/pdt>.

GIBSON, J., K. BEEGLE, J. DE WEERDT et J. FRIEDMAN (2013), "What does Variation in Survey Design Reveal about the Nature of Measurement Errors in Household Consumption?", *World Bank Policy Research Working Paper*, No.6372, février.

GIVORD, P. (2010), « Méthodes économétriques pour l'évaluation de politiques publiques », *Document de travail de l'INSEE*, G 2010 / 08, juin.

GIZ (2012), "Measuring Impacts using Experimental Designs and their Application at GIZ", juin.

GURGAND, M. et M. VALDENNAIRE (2012), « Le fonds d'expérimentation pour la jeunesse et les politiques éducatives : premier retour d'expérience », *Éducation et formations*, n° 81, mars.

IDB (2012), "DEO, Development Effectiveness Overview".

IEG WORLD BANK (2012), "World Bank Group Impact Evaluations – Relevance and Effectiveness", juin

IEG WORLD BANK (2009), "Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System".

IOB (2011a), "Between High Expectations and Reality: An Evaluation of Budget Support in Zambia (2005 – 2010)", *Synthesis Report*, novembre.

IOB (2011b), "Impact Evaluation of Drinking Water Supply and Sanitation Programmes in Rural Benin, the Risk of Vanishing Effects", novembre.

IRD – DIAL (2013), "Protocole d'évaluation d'impact PILAEP », Note interne (non publiée).

JERVEN, M. (2013), *Poor numbers: How we are Misled by African Development Statistics and what to do about it*, Cornell University Press.

JICA (2012), Rapport annuel d'évaluation.

JICA (2010), Rapport annuel d'évaluation.

J-PAL (2011a), "Evaluation: What, Why, When", *Support de formation, J-Pal Europe Executive Education Course*, 12-16 septembre, Bruxelles.

J-PAL (2011b), "The Price is Wrong – Charging Small Fees Dramatically Reduces Access to Important Products for the Poor", *J-PAL Bulletin*, avril.

JOHNSON, S. (2013), "Impact Evaluation of Mobile Banking", Presentation, Impact Evaluation for Financial Inclusion: a Retreat for CGAP Members with Leading Researchers, 10-11 janvier.

KLASEN, S., T. LECHTENFELD, K. MEIER et J. RIECKMANN (2012), "Benefits Trickle Away: the Health Impact of Extending Access to Piped Water and Sanitation in Urban Yemen", *Journal of Development Effectiveness*, volume 4 (4), pp. 537-565, décembre.

KLEEMANN, L. et M. BÖHME (2013), "Happily Uninformed? The Demand for Research among Development Practitioners in Germany", *PEGnet Survey report*.

LABROUSSE, A. (2010), « 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, 1^{er} semestre / Printemps.

LEGOVINI, A. (2010), "Development Impact Evaluation Initiative: a World Bank Strategic Approach to Enhance Developmental Effectiveness", avril.

LELAND, J. (2011), "Development from the Bottom of the Pyramid - An Analysis of the Development Innovations Ventures Program", WISE Program – ASME, Été.

LEVINE, D., R. POLIMENI et I. RAMAGE (2012), "Insurance Health or Insuring Wealth? An Experimental Evaluation of Health Insurance in Cambodia", *Ex-Post Impact Analyses Series*, AFD, No.8, mars.

MCC (2012), "MCC's First Impact Evaluations: Farmer Training Activities in Five Countries", *Issue Brief*, octobre.

McKENZIE, D. (2011a), "How can we Learn Whether Firm Policies are Working in Africa? Challenges (and Solutions?) for Experiments and Structural Models", *Journal of African Economics*, 20 (4), pp. 600-625.

McKENZIE, D. (2011b), "In Defense of Single Project Evaluations: a Response to Ravallion", *World Bank Development Impact Blog*, 26 mai.

MELIA, S. (2014), "Do Randomized Control Trials Offer a Solution to 'Low Quality' Transport Research?", *Working Paper, University of the West of England, Bristol*.

MORRA IMAS, L. et R. RIST (2009), "The Road to Results: Designing and Conducting Effective Development Evaluations", Interactive textbook at www.worldbank.org/r2r.

MOSLEY P. (1987), "Foreign Aid: Its Defense and Reform", *University Press of Kentucky*.

NATIONS UNIES (2013), "World economic situation and prospects".

NONIE (2009), "Impact Evaluations and Development – NONIE Guidance on Impact Evaluation".

OCDE (2013), « Approche méthodologique pour les évaluations de l'aide budgétaire ».

OCDE (2012a), « Nouvelles orientations relatives à la mesure et au suivi par le CAD du financement extérieur du développement », Réunion à haut niveau du CAD, décembre.

OCDE (2012b), "Rural Water and Sanitation – Assessing Impacts", *Evaluation Insights*, No.6, juillet.

OCDE (2010), "Evaluation in Development Agencies, better Aid", *OECD Publishing*.

ODI (2012), "Making Systematic Reviews Work for International Development Research", *Briefing Paper*, No.1, janvier.

ODI and TRIPLELINE CONSULTING (2012), "Process Evaluation of the International Initiative for Impact Evaluation (2008 – 2008)", *rapport final*, septembre.

OLIVIER DE SARDAN, J-P. et V. RIDDE (2012), « Gratuité des soins : une évaluation des politiques publiques », *Afrique Contemporaine*, 2012/3, n° 243.

OZLER, B. (2013), "Defining Conditional Cash Transfer Programs: An Unconditional Mess", *World Bank Development Impact blog*, 13 mai.

PALMER-JONES, R. (2013), "Impact Evaluation, Replication and Ethics", Presentation at the CDI Seminar "Impact, Innovation and Learning: Towards a Research and Practice Agenda for the Future", 26-27 mars.

PARIENTÉ, W. (2008), « Analyse d'impact : l'apport des évaluations aléatoires », *Stateco*, n° 103.

PARIS21 (2012), "Partner Report on Support to Statistics – PRESS 2012 Round", décembre.

PICCIOTTO, R. (2012), "Experimentalism and Development Evaluation: will the Bubble Burst?", *Evaluation*, 18(2), pp. 213-229.

QUENTIN, A. et I. GUÉRIN (2013), « La randomisation à l'épreuve du terrain - L'exemple du projet de microassurance SKY au Cambodge », *Revue Tiers Monde*, 2013 / 1 – n° 213, pp. 179-200.

RALLU, J.-L. (2010), « L'appareil statistique des OMD : un potentiel à développer pour les INS », *Stateco*, n° 105, 99. 19-32.

RAVALLION, M. (2012), "Fighting Poverty One Experiment at a Time: a Review of Abhijit Banerjee and Esther Duflo's *Poor Economics: A radical Rethinking of the Way to Fight Global Poverty*", *Journal of Economic Literature*, 50:1, pp. 103-114.

RAVALLION, M. (2011), "Should you Care about Impact Heterogeneity?", *World Bank Development Impact blog*, 25 septembre.

RIDDE, V. et S. HADDAD (2013), « Pragmatisme et réalisme pour l'évaluation des interventions de santé publique », *Revue d'épidémiologie et de santé publique*, Volume 61, Supplément 2, juin, pp. S95-S106.

ROETMAN, E. (2011), "A can of worms? Implications of Rigorous Impact Evaluations for Development Agencies", *3IE Working Paper*, No.11, mars.

ROSENZWEIG, M. R. (2012), "Thinking Small: a Review of *Poor Economics: A Radical Rethinking of the Way to Fight Global poverty* by Abhijit Banerjee and Esther Duflo", *Journal of Economic Literature*, 50:1, pp. 115-127.

RUBEN, R. (2012), « Dimensionner l'aide au développement : ce que nous enseigne l'évaluation », *Revue d'économie du développement*, 2012 / 4, Volume 26, pp. 95-123.

RUPRAH, I. J. (2008), "You can Get it if you Really Want: Impact Evaluation Experience of the Office of Evaluation and Oversight of the Inter-American Development Bank", *IDS Bulletin*, volume 39, No. 1, pp. 23-35, mars.

SEVERINO, J.-M. et O. RAY (2010), « La fin de l'aide publique au développement : mort et renaissance d'une politique publique globale », *Document de travail du FERDI, Série « Politiques de développement »*, P11, mai.

SFE (2010), « Évaluation des impacts des programmes et services publics », *Les cahiers de la SFE*, n° 6, mai.

SNILSTVEIT B. et H. WADDINGTON (2009), "Effectiveness and Sustainability of Water, Sanitation, and Hygiene Interventions in Combating Diarrhea", *Journal of Development Effectiveness*, Volume 1, Issue 3, pp. 295–335.

STERN, E., N. STAME, J. MAYNE, K. FORSS, R. DAVIES et B. BEFANI (2012), "Broadening the range of designs and methods for impact evaluations", *DFID Working Paper*, No.38, avril.

UNEG (2013), "The Role of Impact Evaluation in UN Agency Evaluation Systems: Guidance on Selecting, Planning and Managing Impact Evaluations", *Guidance Document*, février.

USAID (2012), "Evaluation Policy: Year One – First Annual Report and Plan for 2012 and 2013", Washington, D.C, février.

USAID (2011), "Evaluation, Learning from Experience", *USAID Evaluation Policy*, janvier.

WEYRAUCH, V. et G. LANGOU (2011), "Sound Expectations: from Impact Evaluations to Policy Change", *3IE Working Paper*, No.12, avril.

WHITE, H. (2011), "Five Challenges Facing Impact Evaluation",
<http://nonie2011.org/?q=content/post-2>.

WHITE, H. (2009), "Theory-Based Impact Evaluation: Principles and Practice", *3IE Working Paper*, No.03, juin.

WHITE, H. et D. PHILLIPS (2012), "Addressing Attribution of Cause and Effect in Small and Impact Evaluations: Towards an Integrated Framework", *3IE Working Paper*, No.15, juin.

3IE (2011), "Providing Evidence to Influence Policy", *Annual Report*.

Précédentes publications de la collection

- À SAVOIR N° 1 : La régulation des services d'eau et d'assainissement dans les PED
The Regulation of Water and Sanitation Services in DCs
- À SAVOIR N° 2 : Gestion des dépenses publiques dans les pays en développement
Management of public expenditure in developing countries
- À SAVOIR N° 3 : Vers une gestion concertée des systèmes aquifères transfrontaliers
Towards concerted management of cross-border aquifer systems
- À SAVOIR N° 4 : Les enjeux du développement en Amérique latine
Development issues in Latin America
- À SAVOIR N° 5 : Transition démographique et emploi en Afrique subsaharienne
Demographic transition and employment in Sub-Saharan Africa
- À SAVOIR N° 6 : Les cultures vivrières pluviales en Afrique de l'Ouest et du Centre
Rain-fed food crops in West and Central Africa
- À SAVOIR N° 7 : Les paiements pour services environnementaux
Payments For Ecosystem Services
- À SAVOIR N° 8 : Les accords de libre-échange impliquant des pays en développement ou des pays moins avancés
- À SAVOIR N° 9 : Comment bénéficier du dividende démographique ?
La démographie au centre des trajectoires de développement
How Can We Capitalize on the Demographic Dividend?
Demographics at the Heart of Development Pathways
- À SAVOIR N° 10 : Le risque prix sur les produits alimentaires importés –
Outils de couverture pour l'Afrique
- À SAVOIR N° 11 : La situation foncière en Afrique à l'horizon 2050
- À SAVOIR N° 12 : L'agriculture contractuelle dans les pays en développement –
une revue de littérature
Contract Farming in Developing Countries – A Review
- À SAVOIR N° 13 : Méthodologies d'évaluation économique du patrimoine urbain :
une approche par la soutenabilité
- À SAVOIR N° 14 : Assurer l'accès à la finance agricole
Creating Access to Agricultural Finance – Based on a horizontal study of Cambodia, Mali, Senegal, Tanzania, Thailand and Tunisia

- À SAVOIR N° 15 : *The Governance of Climate Change in Developing Countries*
- À SAVOIR N° 16 : Renforcer la mesure sur la qualité de l'éducation
- À SAVOIR N° 17 : Gérer l'instabilité des prix alimentaires dans les pays en développement
Managing food price instability in developing countries
- À SAVOIR N° 18 : La gestion durable des forêts tropicales –
De l'analyse critique du concept à l'évaluation environnementale
des dispositifs de gestion
- À SAVOIR N° 19 : L'Afrique et les grands émergents
- À SAVOIR N° 20 : *Abolishing user fees for patients in West Africa:
lessons for public policy*
- À SAVOIR N° 21 : Coopérations Sud-Sud et nouveaux acteurs de l'aide
au développement agricole en Afrique de l'Ouest et australe –
Le cas de la Chine et du Brésil
- À SAVOIR N° 22 : L'enseignement privé en Afrique subsaharienne :
enjeux, situations et perspectives de partenariats public-privé
- À SAVOIR N° 23 : Les stocks alimentaires et la régulation de la volatilité des prix
en Afrique
- À SAVOIR N° 24 : Les enjeux du développement en Amérique latine (*Deuxième édition*)
Los desafíos del desarrollo en América Latina (Segunda edición)
Os desafios do desenvolvimento na América Latina (Segunda edição)
- À SAVOIR N° 25 : Tools for what trade? – Analysing the Utilisation of Economic
Instruments and Valuations in Biodiversity Management
- À SAVOIR N° 26 : Gestion du patrimoine urbain et revitalisation des quartiers anciens :
l'éclairage de l'expérience française

Qu'est-ce que l'AFD ?

Établissement public, l'Agence Française de Développement (AFD) agit depuis soixante-dix ans pour combattre la pauvreté et favoriser le développement dans les pays du Sud et dans l'Outre-mer. Elle met en œuvre la politique définie par le Gouvernement français.

Présente sur quatre continents où elle dispose d'un réseau de 70 agences et bureaux de représentation dans le monde, dont 9 dans l'Outre-mer et 1 à Bruxelles, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète : scolarisation, santé maternelle, appui aux agriculteurs et aux petites entreprises, adduction d'eau, préservation de la forêt tropicale, lutte contre le réchauffement climatique...

En 2012, l'AFD a consacré plus de 6,9 milliards d'euros au financement d'actions dans les pays en développement et en faveur de l'Outre-mer. Ils contribueront notamment à la scolarisation de 10 millions d'enfants au niveau primaire et de 3 millions au niveau collège, et à l'amélioration de l'approvisionnement en eau potable pour 1,79 million de personnes. Les projets d'efficacité énergétique sur la même année permettront d'économiser près de 3,6 millions de tonnes d'équivalent CO₂ par an.

www.afd.fr

Les évaluations d'impact dans le domaine du développement

État des lieux et nouveaux enjeux

Les évaluations d'impact, notamment expérimentales, se sont multipliées dans le domaine du développement depuis le début des années 2000, quoique de façon très hétérogène selon les bailleurs. Ces études ont eu le mérite de créer un renouveau de l'intérêt pour la recherche en économie du développement et d'améliorer les standards de qualité de l'évaluation des projets des bailleurs. Elles ont apporté un ensemble de connaissances nouvelles, qui ont eu, dans certains cas, un retentissement considérable et se sont traduites par des décisions politiques fortes. Néanmoins, l'apparition de cette « nouvelle » méthode qui est venue, dans une certaine mesure, bouleverser les pratiques des bailleurs traditionnels et plus généralement des partenaires du développement, ne s'est pas faite sans controverses. Les évaluations d'impact ont ainsi suscité d'intenses débats à la fois dans les milieux académiques et des bailleurs, en termes de méthodes et d'utilisation. Ce document propose de faire un état des lieux de ces débats, notamment des points de consensus atteints au niveau international, et de donner des pistes d'utilisation raisonnée de cet outil. Il aborde également des questions moins débattues et appelle à leur meilleure prise en compte dans la définition des stratégies des bailleurs.

Mots-clés : évaluations d'impact, expérimentation aléatoire contrôlée, aide publique au développement, évaluation des politiques publiques, efficacité de l'aide.

AUTEUR

Stéphanie PAMIES-SUMNER

Département de la Recherche, AFD