UMR 225 IRD - Paris-Dauphine

DOCUMENT DE TRAVAIL DT/2015-01

L'étalon-or des évaluations Randomisées : du discours de la méthode à l'économie politique

Florent BEDECARRATS
Isabelle GUERIN
François ROUBAUD

UMR DIAL 225
Place du Maréchal de Lattre de Tassigny 75775 • Paris Cedex 16 •Tél. (33) 01 44 05 45 42 • Fax (33) 01 44 05 45 45
• 4, rue d'Enghien • 75010 Paris • Tél. (33) 01 53 24 14 50 • Fax (33) 01 53 24 14 51
E-mail : dial@dial.prd.fr • Site : www.dial.ird.fr

L'ETALON-OR DES EVALUATIONS RANDOMISEES : DU DISCOURS DE LA METHODE A L'ECONOMIE POLITIQUE

Florent Bédécarrats
AFD
75012 Paris, France
bedecarratsf@afd.fr

Isabelle Guérin
IRD-CESSMA
75013 Paris, France
isabelle.guerin@ird.fr

François Roubaud
IRD, UMR DIAL, 75010 Paris
PSL, Université Paris-Dauphine,
LEDa, UMR DIAL, 75016 Paris, France
roubaud@dial.prd.fr

Document de travail UMR DIAL
Février 2015

Résumé

La dernière décennie a vu l'émergence d'un nouveau champ de recherche en économie du développement : les méthodes expérimentales d'évaluation d'impacts par assignation aléatoire. Cet article explore le contraste entre d'une part les limites (nombreuses) et la circonscription (très étroite) du champ réel d'application de ces méthodes et d'autre part leur succès, attesté à la fois par leur nombre et leur forte médiatisation. L'analyse suggère que ce contraste est le fruit d'une conjonction économique et politique particulière, émanant de stratégies novatrices de la part des chercheurs de cette nouvelle école, et d'intérêts et de préférences spécifiques provenant à la fois du monde académique et de la communauté des donateurs.

Mots Clés : Evaluation d'impact, méthode expérimentale, Essai randomisé, Méthodologie, Economie politique, Développement.

Abstract

This last decade has seen the emergence of a new field of research in development economics: randomised control trials. This paper explores the contrast between the (many) limitations and (very narrow) real scope of these methods and their success in sheer number and media coverage. Our analysis suggests that the paradox is due to a particular economic and political mix driven by the innovative strategies used by this new school's researchers and by specific interests and preferences in the academic world and the donor community.

Key words: Impact evaluation, Randomized control trial, Experimental method, Methodology, Political economy, Development.

JEL Code : A11, B41, C18, C93, D72, O10

Introduction

La dernière décennie a vu l'émergence d'un nouveau champ de recherche en économie du développement : les méthodes expérimentales d'évaluation d'impacts par assignation aléatoire (*Randomized Control Trial* ; dans la suite du texte RCT). Si le principe des RCT n'est pas une innovation scientifique, son application à grande échelle dans les pays en développement (PED) est en revanche un phénomène sans précédent. Dérivées des sciences médicales, et déjà utilisées depuis les années 1960 pour l'évaluation des politiques publiques dans les pays développés (essentiellement aux États-Unis ; Oakley, 2000), ces méthodes ont été adaptées aux problématiques et aux contextes des pays pauvres. Elles ont connu un succès retentissant, ce qu'atteste leur multiplication. Leurs promoteurs ont rapidement bénéficié d'une véritable consécration internationale. Les meilleures revues d'économie ont ouvert grand leurs portes aux RCT, comme le montre le nombre sans cesse croissant d'articles publiés et dont le numéro spécial sur le microcrédit est l'exemple de plus récent (Banerjee *et alii*, 2015). Rares sont aujourd'hui les formations académiques prétendant "tutoyer l'excellence" qui ne proposent un cursus spécialisé dans ce domaine, comme dans les grandes universités américaines (Harvard, MIT, Yale, etc.) ou ailleurs (London School of Economics, Écoles d'économie de Paris, de Toulouse ou de Marseille, etc.). Rares également les conférences internationales qui ne programment pas de session dédiée aux RCT avec un succès d'assistance réitéré. Rares enfin, les agences d'aides qui n'ont pas créé de département qui leur soit dédié ou qui n'aient engagé ou financé leurs propres RCT.

Sur le plan méthodologique et de la connaissance en économie du développement, les RCT représentent un indiscutable progrès. Néanmoins et en dépit d'un champ d'application restreint (évaluation de projets spécifiques, localement circonscrits et souvent à petite échelle), les RCT sont aujourd'hui labélisées "Gold Standard" de l'évaluation, à l'aune duquel il conviendrait de jauger toute approche alternative. Présentées par ses adeptes comme une véritable révolution copernicienne en économie du développement, on leur attribue en exclusive le qualificatif de "rigoureuses", voire de "scientifiques". Bien au-delà du champ méthodologique, l'ambition de certains défenseurs des RCT les plus médiatiques, par exemple Esther Duflo, est de fournir une liste exhaustive des bonnes et des mauvaises politiques en matière de développement. L'accumulation d'un nombre toujours plus grand d'études d'impact, sur lesquelles il conviendrait de capitaliser afin d'en tirer les enseignements généralisateurs, serait la voie fondatrice de cette ambition universalisante. Plus concrètement, la suprématie revendiquée des RCT dans le champ de l'évaluation quantitative engendre un certain nombre d'effets pervers qui seront discutés ici : disqualification et effet d'éviction des méthodes alternatives, mobilisation toujours plus grande des ressources allouées, positions de rentes. Il existe donc un véritable hiatus entre l'étroitesse du champ d'application et l'hégémonie revendiquée par les promoteurs des RCT les plus en vue. C'est ce paradoxe que nous nous proposons d'explorer dans cet article.

Dans la première partie, nous présentons brièvement les principes fondateurs des RCT et ses avantages théoriques, notamment en regard des autres méthodes d'évaluation existantes. La deuxième partie est consacrée à la discussion et la circonscription du champ réel d'application de ces méthodes. En identifiant leurs limites, notamment lors du passage des conditions expérimentales idéales à leur application sur le terrain, nous tentons d'établir leur périmètre de validité dans l'ensemble plus général des questions de développement. Enfin, la troisième partie s'attache à décrypter, dans une perspective d'économie politique, la façon dont les différents acteurs

interagissent, les jeux et enjeux de pouvoir et au bénéfice de qui ils opèrent. Elle propose d'expliquer pourquoi ces méthodes jouissent d'une légitimité politique et académique qui va bien au-delà de leur réelle portée scientifique. Nous concluons en proposant notre propre vision en matière de méthode d'évaluation d'impact et en dressant quelques pistes de recherche dans le prolongement de cet article.

I.- La montée en puissance d'une méthodologie

Les méthodes expérimentales visent à comparer la situation issue d'un projet (d'un programme ou d'une politique) à celle qui aurait eu cours s'il n'avait pas été mis en place, afin d'en mesurer l'impact net ; c'est-à-dire, une fois purgé de tous les changements advenus par ailleurs. Toute la difficulté est de construire la situation de référence hypothétique (sans projet ; appelée *contrefactuel*), qui par définition n'est jamais observée. La solution proposée par les méthodes expérimentales consiste à sélectionner par tirage aléatoire deux échantillons au sein d'une même population susceptible de bénéficier de l'intervention. Celui-ci ne sera attribué qu'à l'un des groupes, mais les deux feront l'objet d'enquêtes avant et après le projet. Les propriétés statistiques issues de la théorie des sondages garantissent qu'en moyenne, les différences observées entre les bénéficiaires et les non-bénéficiaires peuvent être attribuées au projet. Comme dans toute méthode probabiliste, les résultats sont connus avec une marge d'erreur (l'intervalle de confiance) qui dépend des caractéristiques de l'échantillonnage (taille, méthode, attrition, etc.).

Les méthodes expérimentales cherchent donc à établir formellement une relation causale entre une intervention et un certain nombre de variables de résultats. D'un point de vue scientifique, et en théorie, elles peuvent être légitimement considérées comme l'alternative la plus convaincante pour identifier l'existence et quantifier l'ampleur de l'impact observé. Dans le champ des évaluations quantitatives, elles apparaissent *a priori* plus robustes que les autres méthodes existantes : en l'absence de groupe de contrôle constitué *ex ante*, les approches *avant-après* ou *sans-avec* ne permettent pas de contrôler les variations du contexte ; les méthodes quasi-expérimentales d'appariements – qui apparient bénéficiaires et non-bénéficiaires sur la base de caractéristiques observables communes– lèvent partiellement cette contrainte. Mais en l'absence de sélection aléatoire *ex ante*, elles omettent les caractéristiques inobservables qui pourraient avoir influencé la sélection et qui différencieraient donc le groupe traité du groupe de contrôle (aversion pour le risque, "dynamisme entrepreneurial", insertion dans des réseaux sociaux, etc.). Toujours dans le champ quantitatif, les RCT permettent de surmonter, *a priori*, le défi méthodologique que constitue la démonstration du sens de la causalité, sans reposer sur des hypothèses économétriques complexes et toujours réfutables. Enfin et plus classiquement, elles se distinguent des méthodes qualitatives (études de cas, monographies, entretiens, observation participante) en proposant une mesure chiffrée de l'impact, hors de portée (et de propos) de ces dernières. La méthode expérimentale est devenue à tel point incontournable pour estimer des relations de causalité, qu'initialement opposée aux techniques économétriques, elle en est devenue le "benchmark", comme en atteste le titre du chapitre introductif ("*The Experiment Ideal*") d'un manuel très populaire d'économétrie (Angrist et Pischke, 2009).

Au-delà de ces avantages génériques (et théoriques), la transposition et la multiplication des RCT aux questions de développement sont à saluer pour d'autres raisons. Nous en citerons trois principales.

En premier lieu, elles ont mis le doigt sur un point aveugle des politiques dans les PED, qu'elles soient nationales ou issues de l'aide publique au développement (APD), à savoir leur manque criant d'évaluation quantitative dans le passé. Des sommes colossales ont ainsi été dépensées sans qu'on ait une idée précise de leur efficacité, ouvrant un boulevard à des critiques radicales sur des bases plus idéologiques que scientifiques (Easterly, 2009 ; Moyo, 2009). L'acceptation du principe des évaluations et leur multiplication ne peuvent que concourir à la redevabilité démocratique aussi bien au Sud qu'au Nord (Cling *et alii*, 2003). En second lieu, les RCT ont donné un nouvel élan à la collecte de données d'enquêtes de première main par les économistes du développement. Longtemps cantonnés à la modélisation d'agrégats macroéconomiques issus des grandes bases de données internationales de qualité douteuse, tout particulièrement en Afrique (Jerven, 2013 ; Devaradjan, 2013), les chercheurs peuvent s'appuyer sur la légitimité accordée aux RCT dans le champ du *mainstream* pour opérer un rapprochement nécessaire avec le terrain et les acteurs. Enfin, jusqu'alors, les PED étaient marginalisés par la recherche économique en raison de leur déficit de données de qualité, en particulier longitudinales. La généralisation des RCT permet de placer la recherche économique sur ces pays au niveau des meilleurs standards internationaux. Elle se présente même comme un progrès méthodologique initié au Sud et transféré vers le Nord.

Ces atouts ont conduit à une montée en puissance foudroyante des RCT dans les PED à partir du milieu des années 2000. Le Lateef Jameel Poverty Action Lab (J-PAL) a constitué l'un des promoteurs les plus influents des RCT, animant une vaste campagne de plaidoyer en leur faveur. Il a été fondé en 2003 par Abhijit Banerjee, Esther Duflo, chercheurs au *Massachusetts Institute of Technology* et Sendhil Mullainathan, chercheur à Harvard. Spécialisé uniquement dans les RCT, reconnu comme un label de qualité en la matière, il organise des formations et des échanges de pratiques au travers d'un réseau affiliant notamment 100 professeurs d'universités[1] ainsi que de nombreux chercheurs. Il facilite l'accès aux financements pour mener des études randomisées et stimule la diffusion des résultats, tant dans le champ scientifique qu'à l'attention des dirigeants politiques. Ce laboratoire est étroitement lié à l'ONG IPA (*Innovations for Poverty Action*), une organisation chargée de répliquer à grande échelle les évaluations et les programmes. En mars 2014, dix ans après sa création, le J-PAL affiche s'être engagé dans pas moins de 573 évaluations (achevées ou en cours) dans 61 pays, avec une croissance exponentielle au fil des ans. L'Afrique arrive au premier rang des terrains d'application (avec 176 expériences), loin devant l'Asie du Sud (123, essentiellement en Inde) et l'Amérique latine (108). Les thèmes de prédilection portent sur la microfinance (177), suivi des secteurs sociaux (santé et éducation ; 139 et 129 respectivement), la question de la gouvernance suscitant un intérêt croissant (118). On notera le rôle central joué par Esther Duflo, qui a participé à pas moins de 46 évaluations (dont 14 en cours), dépassée cependant par Dean Karlan, qui en affiche 78 (dont 36 en cours) ! Ces performances ne laissent pas d'interroger sur la proximité des chercheurs au terrain évoquée plus haut (nous y reviendrons).

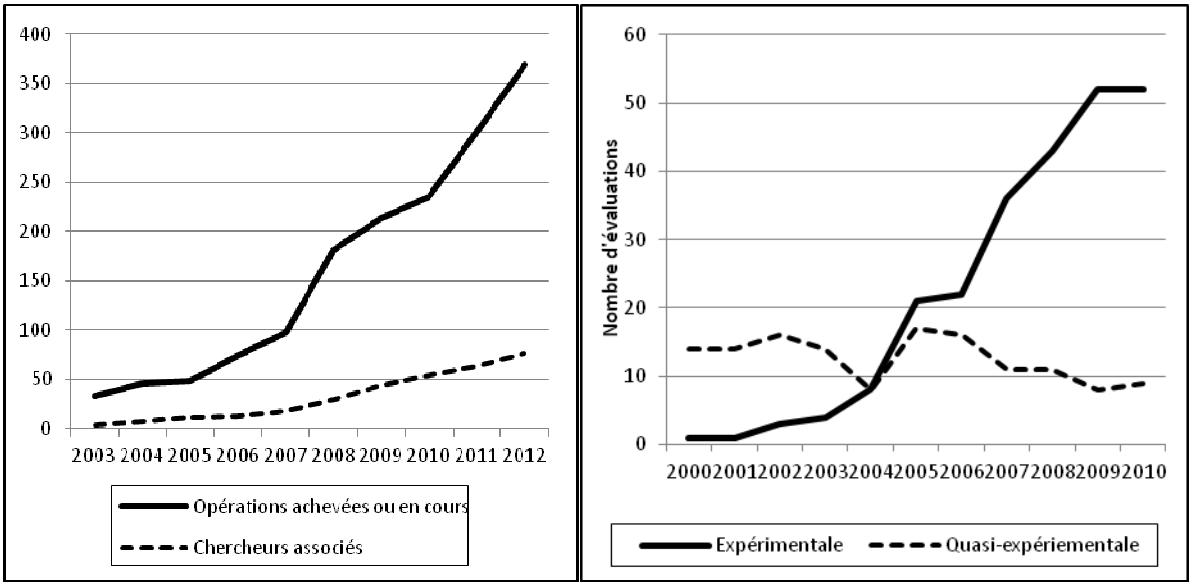
Si la Banque mondiale semble avoir une politique plus nuancée en matière de méthodologie d'évaluation d'impact, en diversifiant les méthodes (néanmoins toujours quantitatives), les RCT en représentent près des deux tiers, avec 64% des 368 évaluations engagées (jusqu'en 2010). Non seulement les RCT tendent à occuper une position de plus en plus dominante, mais elles exercent un effet d'éviction sur les autres approches, comme le montre la figure ci-dessous. Au cours de la période 2000-2004, à peine 20% des évaluations étaient des RCT. Dans les cinq années suivantes, les

1 Informations disponibles sur le site web du J-PAL (http://www.povertyactionlab.org consulté le 26/11/2014).

proportions ont été totalement inversées (76%). Le nombre de RCT est en progression constante, tandis que les évaluations appliquant d'autres méthodes stagnent voire régressent.

Évolution des évaluations d'impact lancées par J-PAL et la Banque mondiale

J-PAL (2003-2012) **Banque mondiale (2000-2010)**



Source : Site J-PAL, février 2013 ; IEG, 2012.

L'évaluation d'impact dans le domaine du développement a connu un succès fulgurant au cours des dernières années, au point de devenir une véritable industrie (plus de 1 000 achevées on en cours, selon White, 2014). Les initiatives locales et internationales ont emboîté le pas (comme 3IE[2] et NONIE). Les méthodes expérimentales ont progressivement phagocyté les autres méthodes et sont aujourd'hui considérées comme le « gold standard » des donateurs comme des chercheurs (Pritchett et Sandefur, 2013). Elles ont essaimé partout dans le monde, ont mobilisé des ressources considérables (en centaines de millions de dollars) et donné lieu à des dizaines d'articles académiques et de nombreux manuels de bonnes pratiques. Face à ce phénomène saisissant (on parle d'un "mouvement pro évaluation d'impact"), il convient de s'interroger sur cet engouement.

Une telle systématisation des RCT est-elle scientifiquement légitime et politiquement souhaitable ? Deux questions font problème : d'une part la revendication de la supériorité intrinsèque des RCT sur toute autre méthode (Duflo *et alii*, 2007) ; d'autre part l'idée que l'accumulation tous azimuts des RCT permettra, par effet de masse, de répondre à toutes les questions de développement, sur "ce qui marche et ce qui ne marche pas" sur des bases incontestables (Duflo et Kremer, 2005).

[2] 3ie (*International Initiative for Impact Evaluation*) est une ONG internationale créée en 2008 et soutenue par de grandes fondations et des bailleurs de fonds, spécialisée dans le promotion de programmes et de politiques de développement « fondés sur les preuves » (*evidence based*) dans les pays en développement. Fin 2013, elle avait financé 150 études d'impact et revues thématiques, pour un montant de 66,6 millions de dollars. Le NONIE (*Network of Networks for Impact Evaluation*) regroupe l'ensemble des réseaux de promotions des études d'impact.

II. La mise en œuvre de la méthode : un champ d'application très circonscrit

Les limites relatives aux RCT sont de trois ordres : des biais pouvant limiter la fiabilité des résultats obtenus (validité interne), la portée restreinte des résultats pour expliquer d'autres situations que le cas particulier étudié (validité externe) et les apories du projet visant à en généraliser l'usage. Nous les déclinerons successivement.

Validité interne

Les tenants des RCT en économie du développement les ont importées du champ médical en faisant abstraction des discussions critiques, des conditions de mises en œuvre et remises en cause dont elles ont fait l'objet en santé publique (Labrousse, 2010 ; Eble *et alii*, 2014). Ils ont également fait l'impasse sur les controverses qui avaient marqué plusieurs décennies de débats en économie du développement.

Les RCT n'échappent pas au « bricolage » habituel des protocoles de recherche, tout particulièrement en sciences sociales. Dans de nombreux cas, les contraintes pratiques et les considérations éthiques nous amènent loin du monde idéal de l'expérimentation de laboratoire. Ces décalages avec la théorie sont susceptibles de remettre en question le principe même du tirage aléatoire, de l'absence de biais de sélection et donc de la supériorité des RCT par rapport à d'autres techniques statistiques et économétriques (Scriven, 2008 ; Labrousse, 2010 ; Deaton, 2010). Pour des raisons d'ordre éthique, il arrive que les partenaires de terrain refusent le tirage aléatoire des bénéficiaires, obligeant les équipes de recherche à opter pour d'autres modes de sélection, par exemple le tirage alphabétique. Or celui-ci peut engendrer des biais, dans la mesure où de nombreuses ressources sont également allouées de manière alphabétique (Deaton, 2010). Il arrive également que les responsables des projets dont on évalue l'impact imposent des protocoles de sélection spécifique afin de ne pas perturber leurs méthodes d'interventions. Citons l'exemple d'un projet de microassurance santé au Cambodge, où les participants ont été tirés au sort lors d'une réunion de village. Or il y a de fortes chances pour que les participants aux réunions villageoises soient plus disponibles, plus curieux et ouverts à l'innovation, plus proches du leader du village, plus insérés socialement, davantage malades, etc., et curieusement, cette question n'est pas discutée (Quentin et Guérin, 2013). Dans certains cas donc, on peut douter de la neutralité des protocoles de sélection et de leur capacité réelle à éliminer les différences de « caractéristiques inobservables » entre groupe traité et groupe de contrôle, que les RCT sont supposées gommer.

Il arrive également que les protocoles de tirage aléatoire se heurtent à la résistance ou au désintérêt des bénéficiaires. C'est notamment le cas lorsque l'intervention suppose un enrôlement volontaire, ou qu'elle entraine des coûts pour les participants. Il est alors extrêmement difficile de s'assurer que les personnes sélectionnées aléatoirement y souscriront ou que ceux qui ont été exclus par tirage au sort n'y auront pas accès par d'autres voies[3]. Or, lorsque l'adhésion des bénéficiaires est plus faible que prévu et incompatible avec les exigences de précision statistique, les équipes de recherche n'hésitent pas à forcer les conditions de participation. Dans deux études menées récemment, l'une sur la microassurance citée plus haut, l'autre sur le microcrédit rural au Maroc, des mesures spécifiques de sensibilisation, d'incitation auprès du personnel de terrain et de rabais sur les prix ont été mises en place à l'instigation des chercheurs (Bernard *et alii*, 2012 ; Quentin et Guérin, 2013 ;

[3] Voir par exemple Heckman *et alii* (1998) ; Rosholm et Skipper (2009).

Morvant-Roux *et alii*, 2014). Les programmes évalués s'éloignent alors sensiblement du programme « normal ».

Le principe de la randomisation suppose que le groupe de contrôle (une population et/ou un territoire) reste exempt de toute intervention tout au long de l'étude. Mais lorsqu'il s'agit de secteurs dynamiques ou concurrentiels (par exemple le microcrédit), il semble illusoire d'espérer des groupes de contrôle qu'ils restent à l'écart de toute intervention et qu'ils n'en subissent pas les effets, contrairement à ce que prévoit le protocole de recherche. Les phénomènes d'attrition sont également problématiques. Le fait que les participants abandonnent le programme (le traitement) au fil du temps est fréquent. Egalement inévitable, la déperdition d'individus échantillonnés, d'une vague d'enquête à l'autre, et ce pour diverses raisons (mortalité, migration, refus de répondre, etc.), d'ailleurs pas nécessairement semblables entre groupes de traitement et de contrôle. Or si l'attrition est plus marquée chez certains types de participants/enquêtés, cela réintroduit *a posteriori* des biais de sélection. Même lorsque les chercheurs essaient de tenir compte des effets de l'attrition dans leurs analyses, ce qui est loin d'être systématique (Elbe *et alii*, 2014), il est très difficile de corriger les biais qui en résultent[4]. En dehors de cas bien spécifiques, les effets de contamination et les externalités de tous ordres, comme l'attrition restent éminemment difficiles à prendre en compte ; et ce, malgré les tentatives des spécialistes des RCT pour mieux spécifier les contours des différentes populations pour lesquelles les impacts sont estimés (ATE: *Average Treatment Effect*; ATT: *Average Treatment on the Treated*; LATE: *Local Average Treatment Effect*, etc. ; Duflo *et alii*, 2007)[5].

Pour parer aux distorsions induites par les arrangements locaux (par exemple si un bénéficiaire tiré au sort fait don ou revend sa participation à son voisin), le tirage aléatoire suppose aussi des précautions multiples, une collaboration très étroite des partenaires de terrain, ainsi qu'une formation et une motivation adéquate des enquêteurs. Plus généralement, la lourdeur des protocoles de recherche (mode de tirage au sort complexes, mais aussi enquêtes à passages répétés, échantillons de plusieurs milliers de personnes, homogénéisation des modes d'intervention), leur durée[6] et leur coût entrainent des jeux d'acteurs, tant avec les opérateurs du projet évalué que les équipes de recherche locales en charge de la conduite des enquêtes, parfois d'autres intervenants qui peuvent être imposés ou suggérés par les bailleurs (par exemple les gouvernements). Une étude sur la mise en œuvre concrète de l'évaluation d'un programme de micro-assurance santé au Cambodge met en évidence la complexité de ces jeux d'acteurs et surtout les compromis et les ajustements multiples qui en résultent, tant au niveau de l'échantillonnage, des objectifs de la recherche, de l'élaboration des questionnaires, que de l'interprétation des résultats, autant de modifications du protocole initial qui éloignent l'évaluation de ses bonnes propriétés théoriques (Quentin et Guérin, 2013). Alors que ces bricolages statistiques et institutionnels sont courants et

[4] Pour un exemple de travaux proposant des méthodes pour essayer de corriger les effets d'attrition (ainsi que d'autres biais), voir Duflo *et alii* (2007). Mais ces recommandations ne sont en général pas suivies dans la pratique (Elbe *et alii*, 2014).

[5] En français, ATE : effet moyen du traitement sur la sous-population de tous ceux qui auraient pu participer au traitement, qu'ils y aient effectivement participé ou pas (ITT : intention to treat) ; ATT : effet moyen du traitement sur la sous-population de ceux qui ont effectivement participé (TOT) ; LATE : Effet local moyen du traitement.

[6] Même pour des enquêtes à un horizon court (1 an par exemple), les études peuvent durer 4 à 5 ans compte tenu de la lourdeur des protocoles de préparation (et du nombre d'études que les équipes RCT gèrent).

constituent un tour de main notoire des randomisateurs)[7], ils sont généralement occultés lors de la publication des résultats. Leurs conséquences sur la nature des protocoles et donc la validité interne des études sont pourtant déterminantes. Le cas le plus emblématique est l'évaluation par l'IFPRI du programme *Progresa*. Par un travail minutieux d'exhumation de la documentation technique (mais non publiée) de l'évaluation, Faulkner (2014) montre qu'aussi bien le protocole initial (choix de l'échantillon) que sa mise en œuvre (phénomène d'attrition et de contamination) s'éloignent du cadre théorique des RCT. Par exemple sur le premier aspect, les échantillons initiaux des groupes de traitement et de contrôle ont été choisis dans deux univers différents, et non tirés aléatoirement comme l'impose toute RCT. Ce qui n'empêche pas les publications ultérieures de présenter le protocole d'échantillonnage comme vraiment expérimental. Seul l'oubli progressif de ces défaillances a permis de "vendre" l'évaluation de *Progresa* comme l'exemple fondateur du bienfondé des RCT pour estimer l'impact causal des programmes sociaux dans les PED. Comme le souligne l'auteur, mettre l'accent (ou simplement mentionner) les points faibles de l'évaluation aurait probablement été contreproductif compte tenu de l'enjeu en balance : à savoir la promotion internationale des RCT comme l'instrument à la fois original et le plus pertinent pour évaluer l'impact des programmes de développement ; et aussi dans le contexte mexicain, le maintien du programme suite à l'élection présidentielle en vue à la fin des années 1990, et l'alternance qui se profilait.

Au-delà du fait qu'elles sont souvent passées sous silence, ces multiples entorses aux protocoles théoriques au fondement des RCT ne seraient pas si problématiques, si elles étaient sérieusement pris en compte ; ce qui est loin d'être le cas. Ainsi Eble *et alii* (2014) se sont livrés à une méta-analyse comparant l'ensemble des RCT parus entre 2001 et 2011 dans les 50 revues d'économie les mieux classées au niveau international (54 articles), avec un échantillon de même taille, représentatif des RCT en médecine publiés dans les trois "meilleures" revues ; leurs caractéristiques méthodologiques étant considérées comme le standard en vigueur. Ils montrent que les performances des premières sont significativement inférieures à celles des secondes, quel que soit le critère considéré. Les RCT en économie sont de manière systématique plus souvent soumises au risque de chacune des six sources de biais identifiées dans la littérature médicale[8]. Même si les procédures pour contrôler les biais de performance et de détection (introduction de placebo, procédure en double aveugle) sont souvent plus difficiles à traiter dans le cas des programmes sociaux, il n'en reste pas moins nécessaire de tenter d'en discuter l'importance. Au total, les RCT en économie sont donc plus sujettes à surestimer les effets du traitement étudié, ce qui interroge quant à la robustesse des résultats obtenus. Elles

[7] Voir par exemple les discussions sur le blog des évaluations d'impact de la Banque mondiale (http://blogs.worldbank.org/impactevaluations/). Ainsi, dans son *post* sur *Les montagnes russes de la randomisation*, Goldstein (2011a) décrit les difficultés liées aux contraintes suivantes : discuter avec le plus grand nombre d'acteurs possibles, pour convaincre les opérationnels qu'une RCT est indispensable et négocier le protocole permettant de donner une puissance statistique satisfaisante ; gérer les remplacements au sein des gouvernements pour maintenir l'adhésion du maître d'ouvrage ; ajuster le protocole quand la mise en œuvre va plus vite que prévu ; rattraper le coup quand le tirage en public a été réalisé sur une mauvaise liste, etc. Dans un autre *post*, il évoque la nécessité de s'appuyer sur informateurs privilégiés locaux pour pouvoir approcher les populations locales (Goldstein, 2014), mais sans s'interroger sur les biais induits éventuels. Il s'interroge également sur les effets "contre-placebo" dans le groupe de contrôle (Goldstein, 2011b). Dans la même veine Ozler (2013) discute des difficultés liées au tirage aléatoire ou encore des biais de réponse dans les questionnaires (qui sont néanmoins valables pour n'importe quelle enquête par questionnaire).

[8] A savoir les biais de sélection, de performance (changement de comportement lorsque les sujets de l'évaluation connaissent leur statut vis-à-vis du traitement ; également appelés "effet de Hawthorne" ou "effet de John Henry"), de détection (changement de comportement lorsque les évaluateurs connaissent le statut des sujets), d'attrition, de *reporting* et d'imprécision (voir Elbe *et alii*, 2014).

sont aussi moins scrupuleuses lorsqu'il s'agit de documenter ces biais potentiels et finalement de chercher à les corriger. De plus, les RCT les plus récentes, comme celles publiées dans les trois meilleures revues (vs les 47 autres), ne sont pas de meilleure qualité.

Ces défaillances ne s'expliquent pas par le fait que les risques encourus par la généralisation des traitements seraient plus élevés dans le domaine médical qu'en économie. Dans les deux cas, on fait face à des effets potentiellement létaux, tout particulièrement dans les PED. De plus, la qualité des RCT en économie y est moindre que dans les pays développés, ce qui n'est pas le cas en médecine (Eble *et alii*, 2014). Ces défaillances relèvent plus d'une désinvolture liée à de moindres exigences de qualité scientifique ainsi qu'à des procédures de contrôle par les pairs plus lâches.

Validité externe

Sur le plan de la validité externe, l'étroitesse des questions susceptibles d'être traitées par les RCT questionne les enseignements que l'on peut en tirer au-delà de l'intervention étudiée. Pour des raisons de précision statistique, les RCT requièrent des échantillons de population très importants, même lorsqu'il s'agit d'apporter la réponse à une question basique : l'intervention a-t-elle amélioré la situation des bénéficiaires. Pour cela, elles traitent difficilement la question de l'hétérogénéité des effets, que la plupart ignorent pour se concentrer sur les impacts moyens (Heckman, 1991). Or, les politiques de développement peuvent avoir des effets très contrastés en fonction des caractéristiques disparates de leurs bénéficiaires ou de la manière dont elles sont localement mises en œuvre[9]. Le projet fonctionne-t-il mieux dans certains contextes que dans d'autres ? Qu'il s'agisse de l'amélioration des services ou de la réplication à plus large échelle, cette question est pourtant essentielle. Cette difficulté est également liée à la croyance selon laquelle un impact moyen suffirait à capter la complexité des effets (Ravallion, 2009 ; DFID, 2012). Dans le domaine de la microfinance par exemple, un impact moyen n'a guère de sens compte tenu de la diversité de l'offre liée aux modalités des services, mais aussi des demandes pour ce type de services en fonction des usagers (Bouquet *et alii*, 2009 ; Mosley et Hulme, 1998) et des territoires d'implantation (Morvant *et alii*, 2014). Cette difficulté est liée à des contraintes techniques et *in fine* financières. Etudier une diversité d'effets avec une puissance statistique suffisante suppose des échantillons – et donc des coûts –, encore plus considérables.

Afin de maîtriser les paramètres de l'expérimentation, les RCT sont souvent très limitées dans l'espace – ne sont ciblées que des zones spécifiques – et dans le temps. Pour des raisons de coût, mais aussi d'attrition, l'horizon auquel les RCT évaluent l'impact d'une intervention dépasse rarement deux ans. Certaines se cantonnent même à un an, notamment lorsque l'enrôlement (*take up*) est trop faible et que les bénéficiaires sont nombreux à faire défection. Cette contrainte n'est pas ignorée par les randomisateurs[10], et certaines évaluations parviennent à s'étaler sur de plus longues périodes, mais cela reste néanmoins exceptionnel[11]. Cette contrainte temporelle amène souvent à se focaliser sur des résultats intermédiaires et non de long terme. Par exemple l'évaluation d'un programme de lutte contre la pauvreté des agriculteurs mesure l'augmentation de l'usage d'engrais et non le revenu des agriculteurs ou leur niveau de pauvreté (Labrousse 2010 ; Eble *et alii*, 2014). L'évaluation d'un programme de micro-assurance santé mesure la diminution de l'endettement des

[9] Pour l'exemple du microcrédit voir par exemple Morvant-Roux *et alii* (2014).
[10] Voir par exemple le *post* de Goldstein (2011c)
[11] Voir par exemple Baird *et alii* (2012) sur les vermifuges.

familles et non leur état de santé (Quentin et Guérin, 2013). On peut également s'interroger sur la sensibilité des effets à l'épreuve du temps (Labrousse, 2010). Pour de multiples raisons, des impacts constatés à court terme peuvent évoluer, voire s'inverser à moyen ou long terme. Dans le domaine du microcrédit par exemple, les effets négatifs de surendettement observés dans plusieurs régions du monde ces dernières années (Guérin *et alii*, 2013) apparaissent souvent suite à plusieurs cycles de microcrédit. Plus généralement, il est assez rare que les trajectoires des projets de développement soient linéaires, ce qui limite fortement la validité des conclusions obtenues après un laps de temps court (Woolcock, 2009).

Nombre de programmes de développement ont des effets d'entraînement, de composition ou d'équilibre général, c'est-à-dire qu'ils affectent une région ou une filière, positivement ou négativement, bien au-delà des seuls bénéficiaires du programme en question. C'est le cas par exemple de programmes d'accès à l'emploi. Ceux-ci peuvent avoir pour effet d'augmenter l'ensemble des revenus du travail (Ravallion, 2009), mais aussi provoquer des effets de saturation ou de substitution. L'exemple typique est celui du microcrédit, où les entreprises créées grâce à ses prêts peuvent provoquer une saturation des marchés locaux, ou réduire la demande des entreprises concurrentes, voire les pousser à la faillite. Ce type d'effet, très fréquent (Bateman, 2010), est très souvent occulté par les études randomisées.

Le passage à l'échelle, d'une intervention ciblée au niveau d'une localité à une région voire à un pays, est loin d'être trivial. Ceci rend problématiques les tentatives de montée en généralité à partir d'un programme circonscrit dans l'espace. La question du "scaling up" n'est pas seulement un problème technique (externalités, contamination, saturation, effet d'équilibre général, etc.) c'est également un problème d'économie politique. Par exemple, Bold *et alii* (2013) ont pu montrer qu'un programme de contractualisation des professeurs au Kenya, qui avait montré un impact positif sur le niveau d'éducation des élèves lorsqu'il était appliqué à petite échelle par une ONG (RCT), n'en avait plus aucun une fois généralisé et mis en œuvre par l'Etat. Basée sur la mise en place d'une RCT très originale (la première du genre à tester des effets organisationnels et d'économie politique) qui tire avantage de l'extension d'un programme au niveau national, l'étude conclut que l'absence d'effet une fois passé à l'échelle s'explique par le changement d'opérateur du projet : des ONG soigneusement sélectionnées et très motivées, d'un côté; des fonctionnaires gouvernementaux et d'organisations liées (syndicats) de l'autre. Ce biais pourrait même être systématique s'il existait une corrélation entre les lieux/personnes/organisations qui acceptent de mettre en œuvre des RCT et les impacts estimés (Pritchett et Sandefur, 2013). L'argument d'Acemoglu (2010) qui souligne les effets d'économie politique des programmes à grande échelle de la part des groupes dont les sources de rente seraient menacées par les réformes nous paraît particulièrement déterminant. C'est une question de fond pour la généralisation des résultats de RCT portant sur des politiques conduites localement et qu'on voudrait passer à l'échelle.

Sur le plan de leur portée scientifique, les RCT permettent éventuellement de mesurer certains impacts ou de tester plusieurs modalités d'une intervention, mais elles ne permettent pas d'en analyser les *raisons* ni les *processus* sous-jacents (Ravallion, 2009). Les mécanismes à travers lesquels une intervention donnée parvient à un résultat restent un point aveugle des RCT (Rao et Woolcock, 2003 ; Hulme, 2007), lié en partie à leur absence de contextualisation (Pritchett et Sandefur, 2013). Comprendre les impacts (et pas seulement les mesurer) supposerait d'analyser les phénomènes étudiés dans leur globalité, d'examiner la complexité des liens de causalité, les interactions multiples,

dynamiques et contradictoires entre différentes entités, de manière contextualisée et située. Ceci impliquerait également des analyses à la fois au niveau méso (prise en compte du contexte, de la dimension institutionnelle des actions menées, etc.) et micro (analyse compréhensive des comportements des ménages) ainsi que des liens complexes entre différentes entités et différents niveaux. A partir de deux exemples illustratifs pris dans le champ de l'économie de l'éducation (l'impact de la taille des classes et le rendement de l'éducation), Pritchett et Sandefur (2013) suggèrent qu'il est beaucoup plus pertinent de prendre des décisions politiques dans un contexte donné en s'inspirant d'études non randomisées mais menées dans le même contexte que d'études randomisées menées ailleurs. Montant en généralité, ils s'emploient à démontrer formellement que la revendication de validité externe de l'impact estimé des RCT est nécessairement erronée. Lorsque l'on tient compte des effets de contexte, il existe un arbitrage dans le choix d'une "bonne estimation" (validité interne d'une RCT) obtenue dans un environnement différent de celui où l'on cherche à l'appliquer (une autre zone géographique, ou plus étendue -typiquement nationale), et une "mauvaise estimation" (c'est-à-dire qui ne tient pas compte de la sélection dans le traitement) mais appliqué au bon environnement. Ce qui, par voie de conséquence, a pour effet direct de remettre en question l'idée que les RCT sont la meilleure méthode possible (le gold standard), comme le soulignent les auteurs en question. Considérer ses résultats comme plus rigoureux que ceux obtenus par d'autres méthodes conduit à des recommandations de politique erronées. *In fine*, la mise en œuvre des RCT suppose de multiples conditions, qui ne sont respectées que dans des cas bien précis, qualifiés de programmes « tunnels » (Bernard *et alii*, 2012). Ces derniers se caractérisent par des impacts de court terme, des inputs et outputs clairement identifiés, facilement mesurables, des liens de causalité unidirectionnels (A cause B), linéaires et enfin non soumis à des risques de faible participation de la part des populations visées. L'ensemble de ces conditions exclut un grand nombre de politiques de développement, qui mettent en jeu des combinaisons de mécanismes socioéconomiques et des boucles de rétroaction (effets d'émulation, d'apprentissage des bénéficiaires, d'amélioration de la qualité des programmes, effets d'équilibre général, etc.). Dans les termes de référence d'une étude commanditée sur le sujet, certains responsables du DFID estimaient ainsi le champ d'application des RCT à moins de 5% des interventions de développement (DFID, 2012). S'il convient de ne pas prendre ce chiffre au pied de la lettre, il ne fait aucun doute que les méthodes expérimentales ne sont pas adaptées pour évaluer l'impact de la grande majorité des politiques de développement. Dans leur papier plus formalisé, Sandefur et Pritchett (2013) aboutissent à des conclusions similaires[12].

Reconnaissant que les résultats des RCT restent très liés aux contextes chaque fois spécifiques où elles sont appliquées (période, lieu, modalités d'intervention du projet), certains de leurs promoteurs les plus en vue, notamment Esther Duflo, arguent qu'il convient de les considérer comme des « biens publics mondiaux » et de créer une instance internationale chargée de les multiplier (Savedoff *et alii*, 2006 ; Glennerster, 2012). Celle-ci constituerait ainsi une base de données universelle et jouerait le rôle de « chambre de compensation » apportant des réponses sur tout ce qui marche ou ne marche pas en matière de développement (Duflo et Kremer, 2005 ; Banerjee et Hee, 2008)[13]. Toutefois, de

[12] « *The scope of application of "planning with rigorous evidence" approach to development is vanishing small* » (Sandefur et Pritchett, 2013, p.1)

[13] Sur le site de J-PAL, une page spécifique est consacrée à ce projet généralisateur et décliné en sous-thématiques (gouvernance, santé, etc.). Le nombre de bénéficiaires de programmes conçus à partir d'expérimentations évaluées par J-PAL (202 millions de personnes en janvier 2014 au total) semble être l'indicateur principal du succès de ce projet généralisateur. L'étroitesse du spectre couvert interroge également

par leurs caractéristiques évoquées plus haut, les RCT se focalisent sur de petits dispositifs, relativement simples et facilement actionnables qui ne sauraient, mises bout à bout, retranscrire l'intégralité des enjeux du développement ou fonder une politique sociale. Les limites évoquées plus haut en matière de validité externe rend illusoire la prétention des randomisateurs à proposer un panier de politiques globales sur la base de RCT nécessairement localisées. C'est d'autant plus vrai, qu'il n'y a pas de loi en sciences sociales comme il en existe en sciences dites "dures" (physiques ou naturelles). Il n'existe donc pas de paramètres universels à estimer équivalents aux constantes gravitationnelles, d'Euler, etc. La question d'économie politique de l'extension à échelle (par exemple au niveau national) de politiques évalués dans des conditions expérimentales, et donc la nécessité de mobiliser les institutions publiques structurellement faibles nous paraît particulièrement difficile à résoudre. Ainsi, François Bourguignon, chercheur renommé ayant largement contribué à promouvoir les RCT, considère cette proposition absurde et scientifiquement infondée[14]. Dans ces conditions, la poursuite de ce projet pharaonique est au mieux une fuite en avant, mais plus probablement le fruit d'intérêts qu'il convient d'identifier.

III.- Éléments d'économie politique d'une entreprise scientifique

Comprendre les décalages entre les limites de la méthode et sa très grande légitimité, aussi bien dans le champ académique que politique, suppose de s'interroger sur les rapports de force qui sont en jeu et qui contribuent à forger les préférences collectives en faveur de telle ou telle méthode. L'évaluation d'impact, dont les RCT constituent l'idéal-type, s'est à ce point massifiée qu'il convient aussi de l'appréhender comme une véritable industrie. Comme toute industrie, le marché des évaluations d'impact est la rencontre d'une offre et d'une demande. Cette demande est double : elle provient à la fois de la communauté des donateurs et du monde académique.

Un nouveau *scientific business model*

Du côté des bailleurs de fonds, la seconde moitié des années 1990 et la décennie 2000 marquent la "fin des idéologies" caractéristique de l'ère des ajustements structurels. La fin de la guerre froide a favorisé une émancipation relative de l'aide publique au développement à l'égard du politique. Pendant cette période, la coopération technique et financière ne constituait souvent qu'un registre supplémentaire des rivalités entre blocs. Cette subordination de la coopération à la *realpolitik* a

---

la portée réelle de ce projet généralisateur : formation de la police pour la thématique « économie politique et gouvernance », vermifuges et soutien scolaire pour l'éducation, distribution gratuite de moustiquaires pour la santé. Voir http://www.povertyactionlab.org/scale-ups (consulté le 28 janvier 2015).

[14] François Bourguignon a été directeur de l'École d'économie de Paris entre 2007 et 2013. Précédemment, il a été économiste en chef et premier vice-président de la Banque mondiale à Washington entre 2003 et 2007, période au cours de laquelle il a contribué à la création du DIME. Dans son discours de clôture de la conférence de AFD-EUDN à Paris le 26 mars 2012, il déclarait (extraits) : « *There has been this fashion during the last couple of years on the RCTs. We even heard colleagues, good colleagues, saying that in the field of development, and in the field of development aid, the only fruitful approach from now on was to do random control trials in all possible fields of interventions. And at the end, we'll have a huge map, a huge catalogue saying "This works, this doesn't work". <u>This is crazy!</u> This will never work and, because of that, we absolutely need the other approaches to evaluating policies and programs. The "pure, scientific evidence" on all what is concerned with development is <u>simply completely impossible</u>. We have to live with this imperfect knowledge.* » (souligné par nous)

toutefois été battue en brèche après la chute du Mur de Berlin. Avec la « fin des grands récits », la crise de l'aide, les OMD et le *New Public Management* ont sommé les promoteurs de l'APD d'apporter la preuve de leur utilité (Naudet, 2006).

Le nouveau credo conjugue une focalisation des politiques de développement en faveur de la lutte contre la pauvreté et la mise en avant d'une gestion axée sur les résultats. Ces orientations, formulées dans la *Déclaration de Paris* en 2005, ont été depuis systématiquement réitérées lors de grande conférences internationales sur l'aide publique au développement à Accra en 2008 puis Busan en 2011. La montée en puissance du paradigme de l'*evidence based policies*, qui consiste à fonder toute décision publique sur des preuves scientifiques, réserve aux savants une légitimité nouvelle dans ces arènes politiques. Les RCT répondent en principe à toutes les conditions requises par ce tournant : empirisme agnostique, simplicité apparente (simple comparaison de moyennes), mobilisation élégante de la théorie mathématique (gage de scientificité), et concentration sur les pauvres (enquêtes auprès des ménages).

Du côté académique, et au premier chef de l'économie, la conjoncture est également favorable à la montée en puissance des RCT : défaite des écoles hétérodoxes centrées sur les structures sociales et les processus de domination, recherche des fondements micros de la macro, primat de la quantification et de l'économie dans le champ des sciences sociales et alignement sur les standards en vigueur au Nord (Berndt, 2014 ; Labrousse 2013). Leur simplicité les rend aisément compréhensible par les décideurs. Elles apparaissent donc comme un vecteur privilégié pour éclairer la décision publique.

Les défenseurs des RCT soulignent sa simplicité et sa propension à être comprise par les décideurs et à les convaincre. L'évaluation du programme *Progresa* au Mexique (Skoufias et Parker, 2001) a constitué un prototype de cette méthode et un cas d'école de sa performativité. Les résultats positifs de cette évaluation ont été mis en avant pour que ce dispositif d'allocations sociales conditionnelles (*Conditional Cash Transfers, CCT*) soit maintenu et massifié, alors que l'alternance politique à la tête de l'État mexicain aurait vraisemblablement entraîné sa suppression. L'exemplarité de cette victoire de la preuve scientifique sur les vicissitudes du politique a fourni un argument efficace à ceux qui plaidaient pour que ces méthodes deviennent le fondement de la décision publique en matière de développement[15]; un argument fortement contesté, comme nous l'avons mentionné plus haut (Faulkner, 2014)

La Banque mondiale a également joué un rôle catalyseur à la fois dans la montée en puissance du paradigme de l'*evidence based policies* et des RCT. Elle a tout d'abord été le théâtre d'un revirement scientifique, des études classiques en (macro) économie du développement dont le département de la recherche de la Banque était le sérail, vers des nouvelles approches empiriques et tournées vers la microéconomie. Ce retournement a été amorcé en 2003 avec la nomination de François Bourguignon au poste d'économiste en chef. En 2005, celui-ci a contribué à la création d'une unité entièrement consacrée à l'évaluation d'impact (le DIME: Development Impact Evaluation Initiative), financée sur les fonds du département de la recherche. Il a aussi commandité une évaluation des activités de ce dernier. Celle-ci a étrillée les travaux scientifiques menés par la Banque au cours de la décennie

[15] Le cas de *Progresa* (ultérieurement appelé *Oportunidades*) est emblématique dans la mesure où il couple un type de politique (les CCT) et un type d'évaluation d'impact (les RCT) censées être des modèles du genre (*success story*), chacun dans leur domaine respectif, et qui se renforcent l'un l'autre.

précédente, au motif qu'ils étaient avant tout « *utilisés à des fins prosélytes en faveur des politiques de la Banque, sans avoir été soumis à un regard suffisamment impartial et sceptique* [et] *une carence criante des garde-fous qui commandent de dissocier la recherche et le plaidoyer*. » (Banerjee *et alii*, 2006, p. 6).

Cette critique fut relayée par un rapport d'un groupe de travail international sur le « fossé de l'évaluation », rassemblant de nombreux chercheurs de renom, dont les principaux promoteurs des RCT (F. Bourguignon, A. Banerjee, E. Duflo, D. Levine, etc.), ainsi que des responsables des principales institutions de développement (CAD, Banque mondiale, Bill et Melinda Gates Foundation, Banques africaine et interaméricaine de développement...). Intitulé « *When will we ever learn* », et publié sous la forme d'un appel-programme par le Center for Global Development (Savedoff *et alii*, 2006), ce rapport a eu un très large écho tant dans la communauté scientifique que parmi les praticiens et les décideurs politiques. Au-delà de son argumentation, ce rapport s'apparente aussi à un plaidoyer *pro domo*, puisqu'il a eu pour effet d'accroître la visibilité et la demande pour les travaux de nombre de ses auteurs, et au premier chef les méthodes expérimentales.

L'élan en faveur des RCT a en outre été conforté par l'émergence d'une nouvelle génération de chercheurs. Ils sont jeunes, issus du sérail des meilleures universités (pour la plupart américaines). Ils ont su trouver la formule du carré magique en combinant excellence académique (légitimité scientifique), effort de séduction en direction du public (visibilité médiatique et légitimité citoyenne) et des bailleurs de fonds (demande solvable), investissement massif dans la formation (offre qualifiée), et modèle d'entreprise performant (rentabilité financière) ; toutes ces qualités se renforçant mutuellement. En multipliant des cours dans les cursus universitaires, mais également en proposant des sessions courtes de formation, en assurant des enseignements classiques (en présentiel) mais également sous des formes nouvelles (MOOC), les randomisateurs se donnent les moyens d'attirer des ressources jeunes, motivées et hautement qualifiées[16]. En s'engageant dans une intense activité de communication et de plaidoyer, à l'aide de toute une série de supports de presse ou para-académiques (*policy brief*, blogs, forums de vulgarisation, *happenings*, etc.), ils montrent l'image avenante de chercheurs ayant accepté de sortir de leur tour d'ivoire. En affichant une posture modeste au plus proche du terrain[17], ils incarnent l'engagement, l'empathie et le désintéressement.

L'étude de la trajectoire et des élites et de leurs réseaux s'avère une perspective puissante pour comprendre l'émergence, le déclin ou la diffusion transnationale de paradigmes scientifiques ou politiques (Dezalay et Garth, 2002). Sans prétendre mener cette entreprise à propos des RCT, qui constituerait un programme de recherche à soi seul, nous nous contenterons d'en tracer les linéaments. La figure d'Esther Duflo est l'exemple le plus illustratif de cette mouvance. Jeune chercheuse franco-américaine, E. Duflo cumule les distinctions académiques, dont la célèbre médaille Bates, qui récompense le "meilleur économiste" de moins de 40 ans et qu'elle a reçue en 2010. Elle a à son actif un nombre impressionnant de publications dans les revues d'économie les

[16] En témoigne le succès des formations en ligne spécifiquement dédiées à ce type de méthode, dispensées sur la plateforme eDX du MIT ; voir par exemple le MOOC intitulé *The Challenges of Global Poverty* et dispensé par Banerjee et Duflo ou encore celui intitulé *Evaluating Social Programs*, par deux de leurs collègues de J-PAL, que plus de 1 000 participants ont suivi en 2013.

[17] On a vu plus haut qu'étant donné le nombre de RCT entreprises simultanément, leur connaissance du terrain posait question.

plus prestigieuses, mais elle vulgarise également ses travaux sous la forme d'ouvrages accessibles au grand public et succès de librairie (voir par exemple Banerjee et Duflo, 2011 ; en français : Duflo, 2010). Depuis 2008, elle figure sur la liste des 100 premiers intellectuels mondiaux du magazine américain *Foreign Policy*. En 2011, c'est le *Time* qui la compte parmi les 100 personnes les plus influentes au monde. Fin 2012, elle a été nommée conseillère du Président Obama sur les questions de « développement global ». En France, elle a été la première titulaire de la toute nouvelle chaire "Savoir contre pauvreté" du Collège de France, créée et financée par l'AFD (Agence Française de Développement). Son nom est régulièrement cité comme candidate potentielle à un prochain prix Nobel d'économie.

Ces jeunes chercheurs de la mouvance RCT se distinguent également dans le mode de gestion de leur activité. En montant des ONG ou des bureaux d'étude spécialisés, ils créent les structures idoines pour recevoir des fonds de toutes origines : publique bien sûr, mais également de fondations, d'entreprises, de mécènes, etc., hors des circuits classiques de financement de la recherche publique. Sur ce plan, ils sont en parfaite adéquation avec les nouvelles sources de financement de l'aide que constituent les fondations privées et les institutions philanthropiques, qui se montrent particulièrement enclines à leur confier des études. En parvenant à créer leurs propres guichets de financement, principalement multilatéraux (l'initiative de la Banque mondiale pour l'évaluation d'impact du développement, l'initiative internationale pour l'évaluation d'impact, la Banque africaine de développement, le Fonds Stratégique pour l'Evaluation d'Impact (SIEF)), mais aussi bilatéraux (la coopération espagnole et britannique) ainsi qu'en provenance de grandes fondations (Rockefeller, Citi, Gates), les randomisateurs ont créé un oligopole sur le marché florissant de l'expérience aléatoire, même si la concurrence est aujourd'hui plus vive compte tenu de l'adoption des méthodes RCT par un nombre grandissant d'équipes de recherche.

La nébuleuse organisée autour de J-PAL, co-dirigée par Esther Duflo, constitue le modèle le plus emblématique et le plus abouti de ce nouveau *scientific business model*. Le laboratoire J-PAL proprement dit, créé par A. Banerjee et E. Duflo[18], est l'un des centres de recherche du département d'économie du Massachusetts Institute of Technology. Cet ancrage institutionnel, au sein d'une des plus prestigieuses universités américaines, ainsi que la notoriété de ses dirigeants, sert à la fois de caution académique et de catalyseur. Au côté de J-PAL, Innovations for Poverty Action (IPA) joue un rôle névralgique. Organisation à but non lucratif, outre sa fonction de communication et de plaidoyer en faveur des RCT[19], elle est chargée d'étendre et de répliquer les expériences aléatoires une fois testées par J-PAL. C'est donc l'articulation des deux institutions qui doit permettre de mener à bien le "projet généralisateur" décrit dans la deuxième partie. Annie Duflo, la sœur d'Esther Duflo, en est le directeur exécutif. Dean Karlan (mentionné plus haut pour ses 78 RCT), professeur à Yale et ex-doctorant des deux initiateurs de J-PAL, est lui le fondateur et membre du bureau des directeurs. Comme Abijit Banerjee est également le conjoint d'Esther Duflo, J-PAL/IPA est non seulement une entreprise globale mais également une affaire de famille. Plus largement, les frontières entre les deux institutions sont poreuses, et de nombreux membres et associés y exercent des responsabilités croisées.

[18] Le troisième fondateur, S. Mullainathan, est actuellement professeur à Harvard.

[19] A la différence de J-PAL, contraint par les réserves académiques de rigueur, IPA peut afficher ses objectifs en usant des recettes directement issues du marketing : "*IPA uses randomized evaluations because they provide the highest quality and most reliable answers to what works and what does not*" (voir le site d'IPA : http://www.poverty-action.org/).

L'industrie des RCT est une entreprise rentable sous tous rapports. Elle est académiquement profitable, et il y a tout à gagner à s'inscrire dans cette mouvance (ou tout à perdre à ne pas en être). Il est aujourd'hui très difficile de publier dans des revues d'économie des articles basés sur d'autres approches. Cet effet d'éviction est aussi lié au fait que les promoteurs les plus influents des RCT sont souvent membres des comités de rédaction des plus grandes revues d'économie et d'économie du développement[20]. Le numéro spécial de l'*American Economic Journal: Applied Economics* consacré aux RCT dans le domaine du microcrédit est illustratif à cet égard. Les trois éditeurs scientifiques du dossier sont membres de J-PAL. En dehors de l'introduction générale, chacun co-signe un article, et deux d'entre eux sont membres du comité éditorial (Banerjee et Karlan). Ester Duflo est à la fois Editeure (fondatrice) de la revue et co-auteure de deux des six articles. Si on ajoute que près de la moitié des auteurs/articles (soit 11 sur 25) sont également membres de J-PAL, et quatre autres sont chercheurs associés ou doctorant, on s'éloigne des principes de contrôle par les pairs qui doivent régir la publication scientifique. Mais les profits ne sont pas uniquement symboliques. Se spécialiser dans les RCT est aussi une excellente façon d'obtenir un poste de chercheur ou d'enseignant, comme en témoignent les modes de recrutement actuels dans la discipline économique. C'est aussi une garantie d'obtention de fonds conséquents pour mener ses propres recherches (dans un contexte de pénurie généralisée) et de générer des rémunérations additionnelles très substantielles, sous forme de consultations ou de participations à des instances de direction[21].

Dans ce contexte, on comprend mieux pourquoi les critiques adressées aux méthodes expérimentales sont particulièrement mal accueillies et rencontrent une très forte résistance de la part de leurs promoteurs[22]. Plusieurs stratégies sont mises en œuvre pour asseoir le monopole et éviter un rééquilibrage au tout RCT. Les méthodes alternatives sont disqualifiées, les RCT s'arrogeant le monopole de la scientificité (Harrison, 2011). Les réflexions critiques sont longtemps restées inaudibles, car cantonnées sur des supports de publication marginalisés. Dans de nombreux cas, les résultats des expériences présentés comme des « découvertes » inédites, ne sont en fait que la reprise de conclusions obtenues par des études antérieures. Le subterfuge procède en deux temps. Il s'agit d'abord de discréditer l'essentiel de la littérature existante, considérée a priori comme non rigoureuse, les RCT étant revendiquées supérieures et appréhendées comme le seul mode reconnu d'administration de la preuve. Ce dénigrement presque rituel conduit à faire des connaissances accumulées dans le passé un champ vierge (*reset*), qui permet, dans un second temps, de faire passer

[20] Comme par exemple et parmi les plus fameuses : *Annual Review of Economics, Journal of Economic Literature, The American Economic Journal: Applied Economics, Review of Economics and Statistics, Journal of Development Economics, Review of Development Studies, Journal of Quantitative Economics, Journal of Economic Perspectives*, etc. Encore une fois D. Karlan apparaît comme le plus boulimique : il est membre dans au moins huit comités (board), soit cinq des revues citées plus haut, auxquelles il faut ajouter *Behavioral Science & Policy*, *Stanford Social Innovation Review* et *Journal of Globalization and Development.*

[21] Rappelons que la rémunération des enseignants dans les universités d'élite américaines est sans commune mesure avec celle des universitaires français. D'ailleurs, le rapport d'évaluation de la recherche à la Banque mondiale cité plus haut considérait que les fonctionnaires de la Banque étaient mal payés en comparaison de leurs homologues académiques, ce qui aurait pour effet de faire fuir les meilleurs talents.

[22] Ainsi à titre anecdotique mais symptomatique, le titre de la conférence EUDN 2012 organisée à Paris par l'AFD et consacrée à l'évaluation, *Malaise dans l'évaluatio*n (AFD, 2012), a été peu apprécié par un certain nombre de randomisateurs, qui ont demandé à ce que le titre soit changé. Il a été d'autant plus mal perçu qu'il écornait la geste hagiographique encore en construction en France, et d'autant plus mal venu qu'il émanait de l'AFD, principal bailleur potentiel (et effectif) des RCT en France sur les questions de développement. C'est d'ailleurs l'AFD qui avait contribué à la création et au financement de la chaire "Savoirs contre pauvreté" au Collège de France, attribuée à son instigation à Esther Duflo en 2009.

tout résultat issu de RCT comme une « découverte » majeure, malgré son caractère redondant[23]. La manipulation est d'autant plus facile à faire admettre que les publications tirées de méthodes non expérimentales ne sont presque jamais citées (Labrousse, 2010 ; Nubukpo, 2012).

La prise de pouvoir par les randomisateurs n'est donc pas seulement une question de méthode. Elle s'étend au contenu et aux thèses qui ont cours sur les grandes questions scientifiques et de politique dans le champ du développement. Quelques exemples pris dans le cadre de la microfinance sont de ce point de vue illustratifs de cette propension à faire passer pour originales, des thèses déjà largement explorées dans la littérature[24].

Prêts collectifs vs prêts individuels. Le décalage entre la littérature et la pratique reflète bien la relative autonomie de la recherche *mainstream* par rapport à la pratique et son étanchéité par rapport à des études pourtant robustes menées par ailleurs. La littérature académique publiée dans de prestigieuses revues internationales s'est éprise de la soi-disant révolution financière du prêt de groupe, alors qu'elle était pratiquée par nombre d'institutions de microfinance depuis les années 1960 (Gentil, 1996 ; Harper et Dichter, 2007). La question de l'efficacité du prêt collectif a été tranchée de manière très pragmatique par les acteurs de terrains qui ont dès le milieu des années 1990 développé le crédit individuel (à commencer par la Grameen Bank). Un pan de la recherche plus opérationnelle a montré dès le début des années 2000 que le prêt de groupe n'est justifié que dans certains contextes qui s'y prêtent socialement et économiquement (réseaux sociaux denses mais pas trop hiérarchiques, faible spécialisation des emprunteurs dans le même secteur, etc.). D'autres travaux, de nature académique, et sur la base d'études quantitatives robustes parviennent à des conclusions similaires (Godquin, 2004 ; Sharma et Zeller, 1997 ; Gonzalez-Vega *et alii*, 1996). Or les travaux expérimentaux ignorent complètement ces différents pans de littérature et n'apportent rien de nouveau, hormis le mode d'administration de la preuve. Par exemple les études de Giné et Karlan (2011) sur les effets de la garantie solidaire vs individuelle sur les impayés sont redondantes par rapport aux dizaines d'études déjà consacrées au sujet et notamment aux trois citées ici. Plus largement, la conclusion de randomisateurs selon laquelle « *le microcrédit n'est peut-être pas un ''miracle'', comme on l'a parfois prétendu, mais il permet effectivement aux ménages d'emprunter, d'investir, et de créer ou d'étendre leurs activités* » (Banerjee *et alii*, 2015), n'est pas très novatrice. On la retrouve dans de nombreuses études d'impact parmi les 170 publiées sur la microfinance (dont une dizaine de RCT) entre 1980 et 2010 (Bédécarrats, 2012).

Les exemples pourraient être multipliés à l'envi. Qu'il s'agisse de la caution solidaire (versus caution individuelle), ou de l'épargne, la littérature préexistante aux RCT interroge le caractère novateur des enseignements de ces dernières. Elles arrivent plus d'une quinzaine d'années après que les innovations financières aient été expérimentées, documentées et largement diffusées. Ce constat

[23] Ce mode opératoire bien connu a déjà été utilisé en économie du développement à propos des institutions. Après avoir exclu du champ de la discipline légitime les courants hétérodoxes (écoles de la régulation et des conventions, néo-institutionnalisme, etc.) dont c'était une des originalités, l'économie *mainstream* s'est réappropriée ce sujet, en faisant passer ses résultats comme des nouveautés. Le même phénomène est actuellement à l'œuvre avec l'économie politique.

[24] Les thèses développées par Banerjee et Duflo (2011) dans leur ouvrage *Poor Economics* méritaient à elles seules d'être passées à ce crible.

contredit le message véhiculé par J-PAL et IPA, qui se présentent avant tout comme des dispositifs de tests stimulant l'innovation en matière de politiques sociales[25].

Alors que tout concourrait pour que s'engage une véritable controverse scientifique (au sens de Callon *et alii*, 2001; Knorr-Cetina, 1982) sur la question des expériences aléatoires, celle-ci n'a toujours pas vu le jour du fait de l'asymétrie de pouvoir des acteurs en présence. Certes le débat existe, mais l'affrontement ne s'exerce pas de manière ouverte. Néanmoins, il convient de noter que la multiplication des critiques conduit progressivement à faire bouger les lignes. Encore une fois le numéro spécial consacré au microcrédit peut être cité en exemple. Ainsi, le numéro est fourni avec les bases de données originales, pour répondre au grief d'opacité et faciliter les méta-analyses. L'introduction générale synthétise les réponses apportées (Banerjee *et alii*, 2015). Un modèle théorique est développé, pour répondre aux critiques d'empirisme agnostique. Les questions du taux de *take up*, de précision des estimateurs ou de l'hétérogénéité du traitement sont reconnues (validité interne). La diversité contextuelle est prise en compte à travers la diversité des terrains, des produits et des institutions abordées dans les six articles (validité externe). Mais le changement de posture reste marginal. En particulier, les deux principes fondateurs demeurent : les RCT constituent la méthode la plus rigoureuse pour mesurer l'impact causal ; le projet généralisateur visant à établir ce qui marche et ce qui ne marche pas est maintenu inchangé. Ainsi, l'argument sur la prise en compte de la diversité des contextes est retourné. La similitude des résultats obtenus dans les six pays, considérés comme un "*échantillon raisonnablement représentatif de l'industrie/mouvement du microcrédit à l'échelle mondiale*" (Banerjee *et alii*, 2015, p. 2) permettrait de lever la critique portée sur la validité externe de la méthode.

Conclusion

Cet article s'est employé à décrire la montée en puissance des méthodes expérimentales dans le champ du développement, au point de devenir l'étalon-or de l'évaluation d'impact. Il a ensuite montré les limites méthodologiques des RCT et l'inanité du projet hégémonique caressé par leurs promoteurs. Enfin, il a proposé quelques éléments d'interprétation pour comprendre les facteurs qui ont présidé à l'instauration de cette nouvelle norme internationale, dans une perspective d'économie politique. De notre point de vue, les expériences aléatoires appliquées au développement représentent un progrès méthodologique. Néanmoins, cette avancée s'accompagne d'une double régression : épistémologique d'abord, les promoteurs des RCT partageant une conception positiviste de la science, aujourd'hui surannée ; politiquement ensuite, par le caractère impérialiste d'une démarche prétendant comprendre tous les mécanismes de développement par cet instrument.

Parmi les extensions possibles de cet article, deux pistes paraissent prometteuses : l'une analytique et l'autre méthodologique. Sur le premier front, notre approche d'économie politique méritait d'être complétée par des travaux s'inscrivant dans le registre de l'histoire ou de la sociologie des sciences. Dans une perspective "latourienne", des recherches portant sur les interactions entre la production scientifique, et les conditions sociales, la personnalité des acteurs ou encore l'architecture institutionnelle pourraient s'appliquer fructueusement à l'industrie des RCT, à ses figures tutélaires

[25] Cf. la description que IPA donne de sa stratégie, à la page : http://www.poverty-action.org/about, consultée le 22 janvier 2015.

et à ses centres de recherche les plus en vue : l'intérêt pour la vie de laboratoire ne devrait pas être seulement réservé aux sciences dites "dures" (Latour, 1999). Plus classiquement, et dans la foulée des travaux de Pierre Bourdieu, il conviendrait d'étudier les mécanismes et les stratégies mises en œuvre par les promoteurs des RCT pour assurer leur domination dans le champ scientifique (Bourdieu, 1997 ; Lebaron, 2000). Car si les randomisateurs cherchent à asseoir leur suprématie au sein de l'économie *mainstream*, ils contribuent également à forger la légende d'une supériorité intrinsèque de cette même économie, en tant que discipline, sur les autres sciences sociales[26], et de l'usage du quantitatif et de la mathématique sur tout autre mode d'administration de la preuve. On peut aussi penser à des approches historiques de trajectoire de randomisateurs, à l'instar de ce qui se fait pour des capitaines d'industrie, des hommes politiques ou encore des scientifiques prestigieux.

Sur le second front, notre propos n'est pas de rejeter les RCT, qui constituent une méthode prometteuse... parmi d'autres ; encore convient-il de les mener dans les règles de l'art, et un alignement sur les bonnes pratiques instaurées dans le domaine médical est une nécessité. Si les RCT restent probablement adaptées et légitimes pour certaines politiques circonscrites de manière précise, il est à la fois nécessaire et possible d'employer d'autres méthodes. Celles-ci adoptent une approche pragmatique, définissant les questions de recherche et les outils méthodologiques nécessaires au cas par cas, en concertation avec les partenaires impliqués (opérateurs de terrain, bailleurs, etc.). Elles s'appuient également sur un pluralisme méthodologique, fondé sur l'interdisciplinarité et reconnaissent la diversité des modes d'administration de la preuve (inférence statistique/analyse compréhensive). Il ne s'agit pas de récuser le formalisme ou la modélisation, mais d'en faire un usage contrôlé. En outre, ces approches ne visent pas l'énoncé de lois universelles, mais cherchent à expliciter des liens de causalité propres à une période et à un contexte précis. Les méthodes qualitatives sont mobilisées pour contextualiser les politiques de développement, élaborer des hypothèses originales, identifier des phénomènes nouveaux ou imprévus, et enfin les analyser dans leur globalité en étudiant la complexité des liens de causalité, les interactions multiples, dynamiques et contradictoires entre différentes entités, de manière située.

Les degrés d'articulation et d'intégration entre méthodes (quantitatives, qualitatives et participatives) peuvent être très variés. Plutôt que de miser systématiquement sur la création de données nouvelles, ces méthodes alternatives privilégient, lorsque cela se justifie, l'usage de données existantes, qu'il s'agisse de statistiques officielles ou de données produites par les partenaires locaux en charge de la mise en œuvre des projets/politiques de développement. Cela permet notamment de réduire les coûts et la lourdeur des protocoles d'enquêtes, dont on a vu plus haut les incidences potentiellement néfastes sur la qualité des recherches menées. Cela permet aussi de renforcer les systèmes d'information des programmes évalués ou les appareils statistiques des pays dans lesquelles ils se situent. Après avoir été emportés par l'effet de mode, certains bailleurs de fonds comme l'AFD (2013) et plus tièdement le DFID (2012) se montrent aujourd'hui plus circonspects quant au champ réel d'application des RCT et à leur capacité à répondre aux questions

[26] Ce combat commun des différents sous-champs du *mainstream* est aussi mené au sein du champ de l'économie afin de conserver le monopole de la légitimité académique. En atteste par exemple en France depuis quelques années, la guerre larvée entre les deux associations professionnelles d'économistes (l'*Association Française de Science Economique* et l'*Association Française d'Economie Politique* ; AFSE vs. AFEP) et les combats féroces autour de la création d'une nouvelle section d'économie au sein du CNU intitulée *Institutions, Economie, Territoires et Sociétés* à côté de la section Sciences Economiques (voir aussi note 23).

qu'ils se posent. Il reste à espérer que ce retour à des positions plus nuancées se traduise par des mesures concrètes de soutien à des méthodes d’évaluation alternatives.

Bibliographie

Acemoglu D. (2010), « Theory, general equilibrium, and political economy in development economics », *Journal of Economic Perspective*, 24(3), pp. 17-32

AFD (2013), *Politique d'évaluation*, AFD, Paris, octobre.

Angrist J.D., Pischke J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton (N.J).

Baird S., Hicks J., Kremer M., Miguel E. (2012), « Worms at Work: Long-run Impacts of Child Health Gains », *Unpublished* (http://scholar.harvard.edu/files/kremer/files/klps-labor_2012-03-23_clean.pdf), [consulté le 23 décembre 2014].

Banerjee A., Deaton A., Lustig N., Rogoff K. (2006), *An Evaluation of World Bank Research, 1998-2005*, Banque mondiale, Washington D.C.

Banerjee A., Duflo E. (2011), *Poor Economics: a Radical Rethinking of the Way to Fight Global Poverty*, Public Affairs, New-York.

Banerjee A., Duflo E., Glennerster R., Kinnan C. (2015), « *The Miracle of Microfinance? Evidence from a Randomized Evaluation* », American Economic Journal: Applied Economics, 7 (1), pp. 22-53.

Banerjee A., He R. (2008), Making Aid Work, in W. Easterly W., « *Reinventing Foreign Aid* », MIT Press.

Banerjee A., Karlan D., Zinman J. (2015) « Six Randomized Evaluations of Microcredit: Introduction and Further Steps », *American Economic Journal: Applied Economics*, 7(1), pp. 1-21.

Bardet F., Cussó R. (2012), « Les essais randomisés contrôlés, révolution des politiques de développement ? Une évaluation par la Banque mondiale de l'*empowerment* au Bangladesh », *Revue Française de Socio-Économie*, 10(2), pp. 175-198.

Barrett C. B., Carter M. R. (2010), « The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections », *Applied Economic Perspectives and Policy*, 32(4), pp. 515–548.

Bateman M., (2010), *Why doesn't Microfinance Work ? The Destructive Rise of Local Neoliberalism*, Zed Books, Londres.

Bédécarrats F., Guérin I, Roubaud F. (2013), « L'étalon-or des évaluations randomisées : du discours de la méthode à l'économie politique », *Sociologie pratique*, 2013/2, No.27, pp.107-122.

Bédécarrats F. (2012), « L'impact de la microfinance : un enjeu politique au prisme de ses controverses scientifiques », *Mondes en développement*, No. 158, pp. 127-142.

Bernard T., Delarue J., Naudet J.-D. (2012), « Impact evaluations: a tool for accountability? Lessons from experience at Agence Française de Développement », *Journal of Development Effectiveness*, 4 (2), pp. 314-327.

Berndt C. (2014), « Behavioral economics, experimentalism and the marketization of development », soumis.

Bold T., Kimenyi M., Mwabu G., Nganga A., Sandefur J., DiClemente R.J., Swartzendruber A.L., Brown J.L., Medeiros M., Diniz D. (2013), **«** Scaling up what works: Experimental evidence on external validity in Kenyan education **»**, *Center for Global Development,* Working Paper 321, March.

Bourdieu P. (1997), « Le champ économique », *Actes de la recherche en sciences sociales*, 119, pp. 48-65.

Callon M., Lascoumes P., Barthe Y. (2001), *Agir dans un monde incertain. Essai sur la démocratie technique*, Le Seuil, collection La couleur des idées, Paris.

Cling J.-P., Razafindrakoto M., Roubaud F., (éd.) (2003), *Les nouvelles stratégies internationales de lutte contre la pauvreté*, Economica/IRD, Paris.

Deaton A. (2009), *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*, National Bureau of Economic Research, Cambridge.

Devaradjan S. (2013), « The Africa's statistical tragedy », *Review of Income and Wealth*, 59, pp. 1-7.

Dezalay Y., Garth B.G. (2002), *La mondialisation des guerres de palis. La restructuration du pouvoir d'Etat en Amérique latine, entre notables du droit et « Chicago Boys »*, Liber, Seuil, Paris.

DFID (2012), *Broadening the range of Designs and Methods for Impact Evaluations. Report of a Study commissioned by the Department for International Development*, DFID Working Paper 38, avril.

Dichter T., Harper M., (éd.) (2007), *What's Wrong with Microfinance ?* Practical Action, Rugby, Warwickshire.

Duflo E. (2010), *Lutter contre la pauvreté*, tomes 1 et 2, Seuil, coll. « La République des idées », Paris.

Duflo E., Kremer M. (2005), « Use of randomization in the evaluation of development effectiveness », in Pitman G. K., Feinstein O. N., et G. K. Ingram (éd.), *Evaluating Development Effectiveness*, World Bank Series on Evaluation and Development, vol. 7, New Brunswick, Transaction Publishers, pp. 205–231.

Duflo E., Glennerster R., Kremer M. (2007), « Using Randomization in Development Economics Research: A Toolkit », in T. P. Schultz and J. A. Strauss (éd.), *Handbook of Development Economics*, vol. 4, Elsevier, pp. 3895–3962.

Durand C., Nordmann C. (2011), « Misère de l'économie du développement », *La Revue des livres* No. 1, sept/oct. (http://www.revuedeslivres.fr/misere-de-leconomie-du-developpement-cedric-durand-et-charlotte-nordmann/)

Easterly W. (2009), *Le fardeau de l'homme blanc : L'échec des politiques occidentales d'aide aux pays pauvres*, Markus Haller.

Elbe A., Boone P., Elbourne D. (2014), « Risk and evidence of bias in randomized controlled trials in economics », Mimeo, Brown University.

Faulkner W. N. (2014), « A critical analysis of a randomized controlled trial evaluation in Mexico: Norm, mistake or exemplar? », *Evaluation*, 20(2), pp. 230-243.

Glennerster R. (2012), « The Power of Evidence: Improving the Effectiveness of Government by Investing in More Rigorous Evaluation », *National Institute Economic Review*, 219(1), pp. R4–R14.

Goldstein M. (2014), "Notes from the field: community entry and seatbelts **»,** Development Impact: News, views, methods, and insights from the world of impact evaluation, (http://blogs.worldbank.org/impactevaluations/notes-field-community-entry-and-seatbelts), [consulté le 23 décembre 2014].

Goldstein M. (2011a), « The impact evaluation roller coaster », Development Impact: News, views, methods, and insights from the world of impact evaluation. (https://blogs.worldbank.org/impactevaluations/the-impact-evaluation-roller-coaster) [consulté le 24 décembre 2014]

Goldstein, M. (2011b), **«** Is it the program or is it participation? Randomization and placebos **»,** Development Impact: News, views, methods, and insights from the world of impact evaluation, (http://blogs.worldbank.org/impactevaluations/is-it-the-program-or-is-it-participation-randomization-and-placebos), [consulté le 23 décembre 2014].

Goldstein, M. (2011c), **«** In the long run... **»,** Development Impact: News, views, methods, and insights from the world of impact evaluation**, (**http://blogs.worldbank.org/impactevaluations/in-the-long-run), [consulté le 23 décembre 2014].

Guérin I., Morvant-Roux S., Villarreal M., (éd.), (2013), *Microfinance, Debt and Over-indebtedness. Juggling with Money*, Routledge, Londres.

Harrison, G. W. (2011), « Randomisation and Its Discontents », *Journal of African Economics*, 20(4), pp. 626–652.

Heckman J.*,* Smith J.*,* Taber C*.* (1998), « Accounting for dropouts in evaluations of social programs », *Review of Economics and Statistics*, 80, pp. 1–14*.*

Heckman J. J. (1991), **«** Randomization and Social Policy Evaluation **»**, *NBER Technical Working Paper No. 107.*

IEG (2012), *World Bank Group Impact Evaluation. Relevance and Effectiveness*, Banque mondiale, juin.

Knorr-Cetina K. D. (1982), « Scientific communities or Transepistemic Arnas of Research ? A Critique of Quasi-Economic Models of Science », *Social Studies of Science*, 12, pp. 101-130.

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(2), pp.2 32.

Latour B. (1999), Pandora's Hope: Essays on the Reality of Science Studies, Harvard University Press, Cambridge, Massachusetts.

Lebaron F. (2000), *La croyance économique. Les économistes entre science et politique*, Seuil, Paris.

Moyo D. (2009), *L'aide fatale, Les ravages d'une aide inutile et de nouvelles solutions pour l'Afrique*, Hachette, Paris.

Morvant-Roux S., Guérin I., Roesch M. Moisseron J.-Y (2014), « Adding value to randomization with qualitative analysis: the case of microcredit in rural Morocco », *World Development*, 56, pp. 302-312

Nubukpo K. (2012), « Esther Duflo, ou 'l'économie expliquée aux pauvres d'esprit' », Blog « L'actualité vue par Kako Kubukpo », *Alternatives Economiques*. (consulté le 11 mars 2013).

Oakley A. (2000), « A Historical Perspective on the Use of Randomized Trials in Social Science Settings », *Crime & Delinquency*, 46(3), pp. 315‑329.

Ozler B. (2012), « When Randomization Goes Wrong... », *Development Impact: News, views, methods, and insights from the world of impact evaluation*. (http://blogs.worldbank.org/impactevaluations/when-randomization-goes-wrong) ([consulté le 23 décembre 2014].

Ozler B. (2013), « Economists have experiments figured out. What's next? (Hint: It's Measurement) **»,** *Development Impact: News, views, methods, and insights from the world of impact evaluation*.
(http://blogs.worldbank.org/impactevaluations/economists-have-experiments-figured-out-what-s-next-hint-it-s-measurement) [consulté le 23 décembre 2014].

Pritchett L., Sandefur J. (2013), « Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix », Center for Global Development Working Paper.

Quentin A., Guérin I. (2013), « La randomisation à l'épreuve du terrain. L'expérience de la micro-assurance au Cambodge », *Revue Tiers Monde*, 1(213), pp. 179-200.

Ravallion M. (2009), « Evaluation in the practice of development, *The World Bank Research Observer*, 24(1), pp. 29–53.

Rodrik D. (2008), « *The new development economics: we shall experiment, but how shall we learn?* », John F. Kennedy School of Government, Harvard University.

Rosholm M., Skipper L. (2009), « Is labour market training a curse for the unemployed? Evidence from a social experiment », *Journal of Applied Economics*, 24, pp. 338–365.

Scriven M. (2008), « A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research », *Journal of MultiDisciplinary Evaluation*, 5(9), pp. 11–24.

Savedoff W. D., R. Levine, Birdsall N. (éd.) (2006), *When Will We Ever Learn? Improving Lives Through Impact Evaluation*, Center for Global Development, Washington D.C..

Shaffer P. (2011), « Against Excessive Rhetoric in Impact Assessment: Overstating the Case for Randomised Controlled Experiments », *Journal of Development Studies*, 47(11), pp. 1619–1635.

White H. (2014), « Current Challenges in Impact Evaluation », *European Journal of Development Research*, 26(1), pp. 18-30.