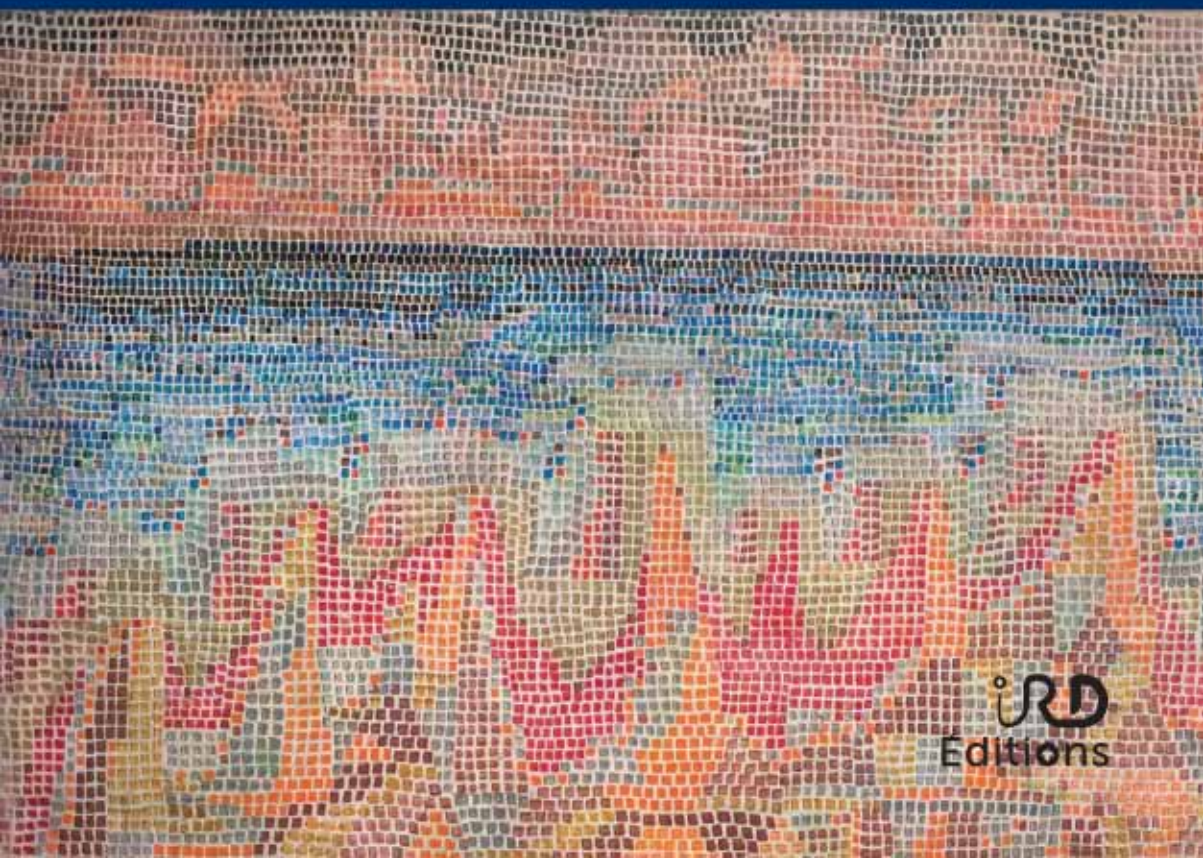
Sous la direction de
F. Bédécarrats, I. Guérin,
F. Roubaud

Expérimentations aléatoires dans le champ du développement

Une perspective critique



IRD Éditions

Expérimentations aléatoires dans le champ du développement

Une perspective critique

Expérimentations aléatoires dans le champ du développement

Une perspective critique

Éditeurs scientifiques
Florent Bédécarrats, Isabelle Guérin et François Roubaud

IRD Éditions
Institut de recherche pour le développement

Collection Synthèses

Marseille, 2022

Ce texte a fait l'objet d'une première édition en anglais :
Bédécarrats F., Guérin I., Roubaud F. (eds), 2020, Randomized Control Trials in the Field of Development : a Critical Perspective, *Oxford, Oxford University Press, 448 p.*
https://global.oup.com/academic/product/randomized-control-trials-in-the-field-of-development-9780198865360

Photo 1re et 4e de couverture
© Ermell – Paul Klee, *Klippen am Meer*, 1931, Musée Lenbachhaus, Munich.
https://upload.wikimedia.org/wikipedia/commons/6/69/Klee_Klippen_am_Meer_1290074.jpg

Coordination éditoriale
IRD/Romain Costa

Préparation éditoriale
Marie-Laure Portal (11)

Mise en page
Desk (53)

Maquette de couverture
IRD/Michelle Saint-Léger

Maquette intérieure
IRD/Pierre Lopez

ISBN papier : 978-2-7099-2947-9
ISSN : 2431-7128
ISBN PDF : 978-2-7099-2948-6
ISBN epub : 978-2-7099-2949-3

Sommaire

Remerciements

L'idée de cet ouvrage collectif est née au cours de l'été 2018 avec la volonté de susciter une controverse scientifique à propos d'une méthode dont l'influence était grandissante et, selon nous, insuffisamment contestée. L'attribution du prix Nobel à Abhijit Banerjee, Esther Duflo et Michael Kremer a eu lieu au moment où nous finalisions le manuscrit : elle rendait la controverse d'autant plus urgente et nécessaire.

Pour mettre en œuvre ce projet, nous n'avons pas bénéficié de « programme » ou de « financement » particulier, ce qui prouve qu'il est encore possible de faire de la recherche de manière libre et désintéressée.

Nous avons néanmoins obtenu un appui financier ponctuel de nos institutions de rattachement – l'Agence française de développement (AFD) et l'Institut de recherche pour le développement (IRD) – pour organiser un séminaire réunissant la plupart des auteurs à Paris en mars 2019. Ce séminaire nous a permis de discuter et débattre, y compris de nos désaccords, puisque l'objet de ce livre n'est pas de converger sur tout, ni de proposer un « prêt à penser » sur la façon d'aborder les RCT, mais d'exposer les termes du débat et les raisons de la controverse. De plus, l'AFD et l'IRD ont accepté de financer à notre demande l'accès ouvert (*open access*) de l'ouvrage en anglais auprès de notre éditeur, la marque d'un engagement ferme en faveur d'une science accessible à tous, notamment au Sud.

Nous remercions donc tous ceux (institutions ou individus) qui ont accepté de faire vivre, chacun à sa mesure, cette controverse trop longtemps éludée : l'AFD et l'IRD pour leur soutien, leurs présidents-directeurs généraux respectifs, qui signent une postface à l'ouvrage, ainsi que Gaël Giraud, alors économiste en chef de l'AFD, qui nous a encouragés à poursuivre l'aventure ; l'ensemble des auteurs de cet ouvrage, dans leur diversité et la pluralité de leurs positions, pour avoir cru en notre proposition, pour avoir joué le jeu de la controverse en donnant le meilleur d'eux-mêmes et pour leur patience ; Lant Pritchett, qui nous a mis en contact avec Gulzar Natarajan et Ila Patnaik, dont les témoignages en fin d'ouvrage apportent l'éclairage du terrain et le point de vue du décideur, un complément nécessaire au regard du chercheur. Il s'est aussi dépensé sans compter pour assurer la promotion du livre ; Angus Deaton, admirable et généreux, qui a toujours répondu présent pour porter à nos côtés la voix de la raison

critique lors des conférences de lancement de l'ouvrage sur les différents continents ; Britta Augsburg, qui, bien que faisant partie de la « communauté » des promoteurs des RCT, a accepté de venir débattre avec nous sur le fond, alors que ceux que nous avons sollicités ont décliné l'invitation ; Diane Bertrand, notre traductrice historique, qui encore une fois a fait des miracles (mais peut-être s'agit-il d'un mirage ?) ; notre éditeur de la version anglaise de l'ouvrage enfin, Oxford University Press, pour s'être engagé, en garant de la liberté académique, à nos côtés, ainsi que les Éditions de l'IRD, qui outre le fait d'avoir accepté de publier l'ouvrage, ont également financé sa traduction en français, avec une mention spéciale au centurion Romain Costa, mandaté pour nous conduire à bon port quoi qu'il en coûte dans la touffeur de l'été 2022.

Gageons que cet ouvrage contribuera à alimenter le brassage des idées, soit la science en action ; à nos futurs lecteurs de le confirmer.

Nous ne saurions conclure ces remerciements à la version française de cet ouvrage sans une pensée émue à Paul Farmer, l'un de nos co-auteurs, tristement décédé en février de cette année 2022. Co-auteur du chapitre 5 sur les expérimentations dans le champ de la santé, Paul était à la fois un scientifique de tout premier plan, y compris dans le champ de l'anthropologie, mais aussi un acteur engagé dans le champ humanitaire, notamment à travers son association Partners in Health qui œuvrait dans plusieurs pays d'Afrique et d'Amérique latine[1]. De par sa position, son regard critique sur le lien entre essais cliniques et RCT dans le champ du développement est particulièrement éclairant. Cette version française de l'ouvrage est aussi dédiée à sa mémoire, dont on espère qu'elle entretiendra l'esprit brillant et critique.

1. Voir par exemple sa nécrologie par Didier Fassin dans le journal Le Monde : https://www.lemonde.fr/disparitions/article/2022/02/28/la-mort-de-l-americain-paul-farmer-medecin-humanitaire-et-anthropologue_6115603_3382.html

Introduction générale

Les controverses sur les expérimentations aléatoires dans le domaine du développement

Épistémologie, éthique et politique

Florent Bédécarrats, Isabelle Guérin et François Roubaud

En octobre 2019, Abhijit Banerjee, Esther Duflo et Michael Kremer se sont vu remettre le 51e prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel. Les trois chercheurs ont été récompensés « pour leur approche expérimentale de la lutte contre la pauvreté dans le monde » et pour avoir « transformé l'économie du développement – domaine qui étudie les causes de la pauvreté dans le monde et les meilleures façons de la combattre –, qui est maintenant un domaine de recherche florissant » (Royal Swedish Academy of Sciences, 2019 : 2). Contrairement à l'expérimentation en laboratoire, l'expérimentation sur le terrain sert à effectuer des tests grandeur nature sur les interventions, les comportements et la prise de décision dans le « monde réel », puis à démontrer des relations causales d'impact (*ibid.* : 3). En conséquence, selon le jury, « nous disposons désormais d'un grand nombre de résultats concrets sur les mécanismes spécifiques à l'origine de la pauvreté et sur les interventions spécifiques visant à l'atténuer » (*ibid.*). Les réalisations des lauréats dans les domaines de la santé, de la scolarisation, de l'égalité des sexes et de la politique, ainsi que du crédit sont de formidables illustrations de leur travail. Ce prix reconnaît le succès d'une méthode de longue date inspirée par le domaine médical, les évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT), et désormais appliquée aux questions de pauvreté et de développement. L'attribution du Nobel d'économie n'a pas vraiment été une surprise. Les RCT ont été introduites dans le domaine du développement dès le début des années 2000 et ont depuis connu un succès

croissant auprès des universitaires, des bailleurs de fonds et des praticiens du développement, à tel point qu'elles sont désormais considérées comme l'étalon-or de l'évaluation des politiques de lutte contre la pauvreté et de la compréhension de ses origines.

S'il y a bien des raisons de saluer l'attribution de ce prix (l'un des trois lauréats est une jeune femme[1] et le prix met en avant la question de la pauvreté et de la collecte de données primaires, longtemps négligée par l'économie du développement), il y a également lieu de s'interroger sur la pertinence et les répercussions du recours accru à cette méthode, que le prix pourrait encore stimuler. Quelle est la portée réelle des RCT ? Ont-elles vraiment « considérablement amélioré notre capacité à lutter contre la pauvreté dans le monde », comme le suggère le jury du prix de la Banque de Suède ? Quels types de questions les RCT sont-elles en mesure de traiter, ou non ? L'explication causale est-elle la seule approche permettant de comprendre la pauvreté, et les RCT parviennent-elles systématiquement à la fournir ? Enfin, et surtout, la prééminence de l'expérimentation dans le domaine de l'économie du développement, telle qu'elle est reconnue et saluée par le jury du prix Nobel, est-elle scientifiquement légitime et politiquement souhaitable ?

Le présent ouvrage propose de répondre à ces questions. L'initiative de ce projet éditorial est issue de la conférence European Development Network (EUDN) sur le *Malaise dans l'évaluation* organisée par l'Agence française de développement (AFD) à Paris en 2012 (AFD, 2012). Lors de cet événement, nous avons assisté à un véritable dialogue de sourds : alors que certaines voix critiques ont exposé les raisons de leurs doutes, ceux que nous appellerons les *randomistas*[2], pour reprendre un terme bien usité dans le milieu et par souci de facilité, exposaient avec assurance leurs convictions et leurs résultats en évitant d'engager le débat sur le fond.

Nous avons donc décidé d'analyser le succès des RCT sous trois angles (Bédécarrats *et al.*, 2013 ; 2019a ; 2021) : en développant des critiques théoriques fondées sur les questions classiques de validité interne et externe (les RCT « par le haut » ; *doing the maths*) ; en portant la critique sur le front empirique : comment les RCT sont menées sur le terrain ? (les RCT « par le bas » ; *doing the cooking*) ; enfin, en analysant l'économie politique des RCT en termes d'offre et de demande (les RCT comme « business » scientifique ; *doing the accounts, both financial and symbolic*). Autant le premier point avait déjà été largement étudié (notre contribution restant donc marginale), autant les deux autres dimensions étaient

1. Le prix Nobel d'économie, décerné pour la première fois en 1969, a depuis été remporté par 84 lauréats. Esther Duflo n'est que la deuxième femme lauréate. Au-delà du prix lui-même, l'économie en tant que science sociale est la discipline la plus marquée par la discrimination à l'égard des femmes (Lundberg et Stearns, 2019).

2. Nous entendons par là les chercheurs qui défendent la supériorité de cette méthode sur toutes les autres. Sur le terme non péjoratif de « *randomista* », voir Ravallion (chap. 1, ce volume) et Ogden (chap. 4, ce volume). Voir aussi Gibson (2019).

relativement inexplorées[3]. Nos propres analyses, issues de l'observation approfondie de deux RCT (microcrédit au Maroc [Morvant-Roux *et al.*, 2014] et micro-assurance au Cambodge [Quentin et Guérin, 2013]), ont été largement corroborées par l'analyse de trois des RCT les plus emblématiques[4]. Ces RCT se sont finalement révélées très discutables, alors qu'elles avaient largement contribué à faire de cette méthode le véritable étalon-or.

Après ces recherches préliminaires, nous avons continué dans deux directions parallèles. Nous avons poursuivi notre travail sur le microcrédit rural au Maroc en menant une réplication. Les résultats ont non seulement corroboré l'hypothèse d'une contradiction entre les RCT en théorie et en pratique, mais ils ont aussi révélé de nouvelles dimensions de cette divergence (Bédécarrats *et al.*, 2019a ; 2021). Étendue à un ensemble de RCT sur le microcrédit, cette contradiction fait l'objet d'un des chapitres de cet ouvrage (chap. 7). Désireux de débattre de la question et de susciter une controverse scientifique, ou du moins une discussion, nous avons ensuite lancé ce projet de co-écriture d'un livre pour ouvrir la question à d'autres disciplines, voix et opinions, y compris des points de vue beaucoup plus positifs sur la méthode que les nôtres. Si certains disent que le débat est fatigant et lassant (Dimova, 2019 ; voir aussi Ogden, chap. 4, ce volume), nous considérons néanmoins qu'il est vital, tant sur le plan scientifique que démocratique, et ce pour des raisons que nous exposons plus avant.

Dans cet ouvrage, nous rassemblons certains des plus grands spécialistes du domaine, issus de divers horizons et disciplines (économie, économétrie, mathématiques, statistiques, économie politique, socio-économie, anthropologie, philosophie, santé globale, épidémiologie et médecine, élaboration des politiques), et examinons les principales faiblesses des RCT dans le domaine du développement, mais aussi quelques-uns de leurs points forts insoupçonnés. Nous prenons des exemples concrets pour expliquer le fonctionnement des RCT, ce qu'elles peuvent

3. Deux conclusions principales sont ressorties de nos analyses. Premièrement, si les RCT représentent un bon moyen d'estimer l'impact causal d'un certain nombre de projets délimités, cela n'est vrai que dans des conditions idéales définies en théorie, mais rarement observées sur le terrain. Dans ces conditions idéales, il est possible que les RCT puissent être utilisées pour quantifier statistiquement les impacts (significativité et ampleur), mais ils ne permettent pas d'identifier les mécanismes par lesquels ces impacts transitent (ce qui est paradoxal pour une méthode qui fait de l'analyse de la causalité son principe fondamental). Deuxièmement, trois des principales affirmations des *randomistas* sont sans fondement : les RCT sont supérieures à toute autre méthode ; la multiplication des RCT peut résoudre le problème de la validité externe, reconnu par tous comme une faiblesse intrinsèque (ce que nous avons qualifié de « projet hégémonique ») ; et les RCT peuvent fournir toutes les réponses sur « ce qui fonctionne et ce qui ne fonctionne pas dans le développement ».

4. La fameuse RCT associée au programme de transferts monétaires conditionnels (*Conditional Cash Transfers* – CCT) au Mexique (*Progresa*, rebaptisé *Oportunidade*s, puis *Prospera*), que beaucoup considèrent comme le catalyseur de la ruée sur les RCT, ainsi que sur les CCT par la même occasion, mais dont la mise en œuvre et donc la validité interne sont contestées (Faulkner, 2014) ; la RCT tout aussi célèbre sur les vers intestinaux au Kenya, réalisée par Miguel et Kremer (2004), dont les résultats ont été contestés par un groupe d'épidémiologistes (Aiken *et al.,* 2015 ; Davey *et al.,* 2015 ; Humphreys, 2015), ce qui est paradoxal car les *randomistas* ont fait des RCT en médecine le fer de lance du mouvement ; et enfin, la RCT sur le recrutement et la supervision des enseignants au Kenya (Duflo *et al.,* 2015), dont Bold *et al.* (2013) ont montré que la transposition à grande échelle *via* une politique nationale mise en œuvre par le gouvernement n'a produit aucun des résultats escomptés.

permettre de mesurer, pourquoi parfois elles échouent, comment elles peuvent être améliorées et pourquoi d'autres méthodes sont à la fois utiles et nécessaires. Nous abordons les questions de méthode, d'épistémologie, d'éthique, de théorie et d'idéologie. En mettant l'accent notamment sur la mise en œuvre des RCT *sur le terrain*, loin des conditions de laboratoire idéales, nous nous distinguons nettement des autres analyses critiques. Cela permet de révéler certaines utilisations et certains effets insoupçonnés de ces RCT, leurs utilisations et leurs fins politiques, mais aussi leur potentiel perturbateur (au sens positif du terme). Nous explorons la vision du monde implicite sur laquelle s'appuient de nombreuses RCT et qu'elles diffusent. Tout en examinant l'écart entre la portée limitée de la méthode et son succès dans le monde entier, nous proposons des domaines d'amélioration, ainsi que des méthodes alternatives. Sans contester la contribution des RCT à la science, nous mettons en garde contre leur prétendue supériorité et les dangers potentiels d'une utilisation inadaptée. Nous soutenons également que le meilleur usage des RCT n'est pas nécessairement celui qui vient immédiatement à l'esprit et que ses partisans promeuvent : comprendre certains comportements plutôt qu'évaluer les interventions.

Si le principe des RCT en science remonte à plus d'un siècle – leur utilisation dans le cadre du développement international est appelée la quatrième vague (Jamison, 2017) –, leur utilisation à grande échelle dans les pays en développement est sans précédent (Ravallion, chap. 1, ce volume). Les RCT représentent une avancée indéniable pour l'économie du développement, et ce pour plusieurs raisons. Elles offrent une solution (parmi d'autres) à l'épineuse question de l'attribution causale (comment isoler l'effet d'une intervention de tous les changements qui se sont produits en même temps). Elles accordent une place centrale à la question de l'évaluation de l'aide et à la nécessité de rendre compte de l'aide. Les RCT donnent un nouvel élan à la collecte de données d'enquête de première main par les économistes du développement. Enfin, par le passé, les pays du Sud 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. La nouvelle vague de RCT dans le domaine du développement se présente même comme un progrès méthodologique commencé au Sud et transféré vers le Nord (Bédécarrats *et al.*, 2019b).

Néanmoins et en dépit d'un champ d'application restreint (détaillé ci-dessous et tout au long de l'ouvrage), les RCT sont aujourd'hui labélisées étalon-or de l'évaluation, à l'aune duquel il conviendrait de jauger toute approche alternative. Cette suprématie est par ailleurs susceptible d'être renforcée par l'attribution du prix de la Banque de Suède. Présentées par leurs adeptes comme une véritable révolution copernicienne en économie du développement[5], on leur attribue en

5. « De la même manière que les évaluations randomisées ont révolutionné la médecine au xx^e^ siècle, elles ont le potentiel de révolutionner les politiques sociales au xxi^e^ » (Duflo, Glennerster et Kremer, 2004 : 29).

exclusive le qualificatif de « rigoureuses », voire de « scientifiques » (voir Ravallion, chap. 1, ce volume). Bien au-delà du champ méthodologique, l'ambition de certains défenseurs des RCT les plus médiatiques est de fournir une liste exhaustive des bonnes et des mauvaises politiques en matière de développement (Labrousse, chap. 8, ce volume). L'objectif invoqué est d'accumuler un nombre toujours plus grand d'études d'impact afin d'en tirer les enseignements généralisateurs. Il est cependant évident que la suprématie revendiquée des RCT en matière d'évaluation engendre un certain nombre d'effets pervers. Citons par exemple la disqualification et l'effet d'éviction des méthodes alternatives, la mobilisation toujours plus grande des ressources allouées, les positions de rentes, et la légitimation d'une vision spécifique et étroite du « développement » (ce que Lant Pritchett appelle dans le chap. 2 de ce volume le « développement fétichiste », *kinky development*). À cela, s'ajoute la disqualification des projets et des politiques de développement qui ne respectent pas les contraintes exigées par les protocoles de randomisation (Ravallion, chap. 1, ce volume ; Garchitorena *et al.*, chap. 5, ce volume ; Patnaik, entretiens, ce volume ; voir aussi Adams, 2016).

Nous ne sommes évidemment pas les premiers à formuler des critiques et de nombreuses voix se sont déjà élevées[6]. Les critiques de James Heckman et Angus Deaton (Deaton, prologue, ce volume ; Deaton, 2010a ; Deaton et Cartwright, 2018 ; Heckman, 1992 et chap. 12, ce volume) sont d'autant plus percutantes que tous deux ont également reçu le prix Nobel d'économie (Deaton en 2015 et Heckman en 2000). Si ces critiques sont désormais plus fréquemment reconnues par les membres du mouvement pro-RCT (Ogden, chap. 4, ce volume),

6. Voir par exemple Heckman, 1992 ; Rodrik, 2009 ; Barrett et Carter, 2010 ; Deaton, 2010a ; Harrison, 2011 ; Pritchett et Sandefur, 2015 ; Deaton et Cartwright, 2018. Plusieurs ouvrages ont également contribué à cette discussion. Le premier, édité par Cohen et Easterly (2010), a mis le feu aux poudres et a déclenché la controverse naissante. Il ne contient qu'un seul chapitre consacré spécifiquement à ce sujet, avec un débat passionnant, bien que bref, entre Banerjee, Rodrik, Mulathain et Ravallion. Les autres chapitres abordent principalement la question de savoir si les politiques de développement fonctionnent et si oui, lesquelles, mais la question des RCT est présente en filigrane tout au long du livre. Le livre de Ogden (2017) est le plus récent et porte essentiellement sur les RCT. Il se présente sous la forme de 20 entretiens avec des acteurs de premier plan dans ce domaine. Quatorze d'entre eux sont des figures actives du mouvement pro-RCT et quatre autres y sont plus modérément impliqués. Seules deux personnes se montrent critiques (Angus Deaton et Lant Pritchett). Malgré leur notoriété, leurs contributions sont assez courtes et copieusement réinterprétées et critiquées par les autres contributeurs. Troisièmement, le livre édité par Teele (2014) est plus détaillé et plus nuancé. Il apporte une contribution majeure à notre compréhension du sujet, en comparant notamment des RCT menées dans des pays du Nord et du Sud par des politologues et des économistes. Néanmoins, il reste centré sur des considérations méthodologiques et épistémologiques. De nombreuses contributions ne font que reprendre des articles aujourd'hui dépassés, publiés ailleurs dans les années 2000, avant que les RCT ne connaissent un véritable essor. Dix ans plus tard, le présent livre actualise les ouvrages précédents en s'appuyant sur la littérature la plus récente et en adoptant une vision plus large, à la fois en termes d'angles disciplinaires et d'enjeux. Enfin, à la fin 2019, alors que nous finalisions notre manuscrit, la revue *World Development* a publié un numéro spécial sur les RCT dans le domaine du développement (paru début 2020). Profitant de l'attribution du prix Nobel d'économie à Banerjee, Duflo et Kremer, ce numéro spécial rassemble un peu plus de 50 courts articles (d'une ou deux pages) rédigés par un large éventail d'auteurs. Compte tenu du format condensé de ces contributions, elles ne peuvent évidemment pas fournir une analyse approfondie. Outre l'examen d'un large éventail de positions concernant les RCT (ce qui a bien fonctionné et ce qui n'a pas fonctionné), ce numéro spécial présente toutefois l'intérêt de proposer des pistes de recherche futures.

la question n'a toutefois jamais fait l'objet d'une véritable controverse scientifique. À défaut de cette controverse (les *randomistas* les plus éminents ont décliné notre invitation), ce livre instaure un dialogue entre les approches, les disciplines, les différents secteurs d'intervention, et, en définitive, entre les différents points de vue sur le rôle des RCT et leur potentiel.

Certains auteurs du livre considèrent que l'engouement pour les RCT constitue une « folie » (Pritchett, chap. 2, ce volume), que leur supériorité n'est autre qu'une « fiction » (Labrousse, chap. 8, ce volume), et que ce sont des « outils inefficaces en matière de redevabilité et d'apprentissage des organisations » et pas à proprement parler des évaluations (Picciotto, chap. 9, ce volume). D'autres considèrent qu'elles ont leur place dans la palette des méthodes d'évaluation, mais que leur supériorité autoproclamée relève « plus de la foi que de la science » et que, dans certaines situations et pour certaines questions, les études non expérimentales (*observational studies*) se révèlent nettement plus appropriées (Ravallion, chap. 1, ce volume). C'est ce que montrent également les analyses sectorielles portant sur la santé (Garchitorena *et al.*, chap. 5, ce volume), l'assainissement rural (Spears, Ban et Cumming, chap. 6, ce volume), le microcrédit (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume) et la gouvernance (Natarajan, entretiens, ce volume).

Un point de vue plus optimiste laisse entendre que les RCT ont pris en compte les critiques et que, sous leur forme actuelle, elles offrent de véritables réponses à de nombreuses questions de développement (Ogden, chap. 4, ce volume). Selon Jonathan Morduch, les RCT sont utiles non pas tant pour « évaluer », mais pour « explorer » les comportements en recourant à des manipulations des structures de prix, des contrats, des méthodes pédagogiques, etc. Les chercheurs peuvent exploiter la perturbation générée par les protocoles randomisés pour observer *in situ* les changements dans les interventions et les comportements, étudier leurs répercussions et en tirer des conclusions opérationnelles (Morduch, chap. 3, ce volume).

D'autres demandent qu'elles soient améliorées tant du point de vue éthique, lequel demeure un angle mort pour les protocoles d'enquête en économie du développement (Abramowicz et Szafarz, chap. 10, ce volume), que du point de vue de l'explication causale, qu'il soit question de mieux utiliser les *a priori* (*priors*) (Vivalt, chap. 11, ce volume) ou des phénomènes de non-respect du protocole (*non compliance*) comme révélateurs des préférences des populations ciblées (Heckman, chap. 12, ce volume).

L'objectif de cette introduction, qui reflète uniquement le point de vue des éditeurs, n'est pas de réconcilier les auteurs ou de trouver un compromis, mais de donner au lecteur une image plus claire des enjeux du débat. Dans la première partie, nous exposons en détail les arguments épistémologiques, politiques et éthiques qui le sous-tendent. Dans la deuxième partie, nous nous efforçons de définir les politiques et les projets de développement qui pourraient se prêter aux spécificités des RCT. La troisième partie revient sur l'idée d'une controverse scientifique, que nous appelons de nos vœux, mais qui n'a malheureusement

pas encore eu lieu, en examinant les raisons de son inexistence. La conclusion propose des moyens d'améliorer les RCT, ainsi que des alternatives méthodologiques.

Les arguments du débat : épistémologie, politique et éthique

Nous ne reviendrons pas ici sur toutes les critiques formulées à l'égard des RCT, qui sont déjà énumérées dans différents chapitres (Ravallion, chap. 1, ce volume ; Ogden, chap. 4, ce volume ; voir aussi Bédécarrats *et al.*, 2019b), car nous pensons qu'il est plus utile de se pencher sur les dissensions épistémologiques, politiques et éthiques qui sous-tendent – souvent implicitement – bon nombre des divergences d'opinions concernant les RCT.

Loin d'être purement techniques, les débats autour des RCT renvoient à des conceptions de la connaissance et du savoir différentes et souvent bien difficiles à concilier. La recherche en sciences sociales sur les interactions humaines est-elle perçue de manière scientiste[7] (Putnam, 2009), comme la recherche de la réponse ultime et universelle à un problème donné, ou comme un processus d'apprentissage continu pour trouver des réponses raisonnables, limitées dans le temps et l'espace, compte tenu de la diversité des connaissances, y compris celles des populations visées par le développement ? Les chiffres, les méthodes statistiques et économétriques appliquées aux sciences sociales ne sont-elles pour nous que des instruments et des techniques, fruits d'un progrès scientifique linéaire ? Ou les considérons-nous aussi comme une construction sociale et politique érigée par des conventions quelque peu arbitraires, inextricablement liées à une certaine conception de l'État et des politiques publiques, du marché, du pouvoir et de l'action collective (Desrosières, 2013b), qui façonnent en partie le monde qu'elles cherchent à représenter, à comprendre et à conseiller (MacKenzie, Muniesa et Siu, 2007) ? Cette deuxième acception de la connaissance ne rejette pas les preuves scientifiques, mais préconise leur ancrage dans des contextes sociaux et politiques particuliers. Par ailleurs, elle différencie clairement les connaissances scientifiques de la prise de décision politique, cette dernière impliquant de se référer à des valeurs pour choisir entre différentes options et évaluer leurs conséquences sociales, économiques et politiques (Drèze, 2018a).

Les divergences d'opinions autour des RCT sont également fondées sur des notions différentes du développement, de la pauvreté et, plus largement, de la

7. Par scientisme, nous entendons l'idée que la science expérimentale est la seule source fiable de connaissances sur le monde et qu'elle est le meilleur moyen d'organiser l'humanité pour résoudre tous ses problèmes les plus urgents. L'expérimentation prétend se passer de tout raisonnement métaphysique, philosophique, éthique ou esthétique.

politique, considérée comme une conception du monde dans laquelle nous vivons et que nous nous efforçons d'atteindre. Le monde est-il un agrégat d'individus à la recherche d'autonomie ou bien un système complexe fait de dialectique, d'interactions multiples, de rétroactions et d'effets systémiques entre des êtres sociaux interdépendants et souhaitant le rester ? Devrions-nous considérer les « causes de la pauvreté comme un manque ou un besoin de ressources pertinentes ou comme un processus actif d'appauvrissement ou de perpétuation de la pauvreté » (SHAFFER, 2015 : 154) ? Si l'on appréhende les causes de la pauvreté sous l'angle des manques, il convient de mettre en place des politiques visant à « faire la différence » (pour faire face aux déficits en matière de santé, d'éducation, de nutrition, d'eau et d'assainissement, de crédit, etc.) ; et pour comprendre les effets de ces politiques, il faut un contrefactuel pour pouvoir isoler cette différence et attribuer l'impact à la politique en question. En revanche, si l'on conçoit la causalité de la pauvreté en termes de processus et de relations sociales, il convient d'adopter des politiques macro-économiques et structurelles (taux de change, politiques de contrôle des capitaux, mesures de protection sociale, etc.) et, pour comprendre les effets de ces mesures, il faut adopter une « approche fondée sur les mécanismes » qui explore la diversité et la complexité des processus causaux à l'origine de ces effets (SHAFFER, 2015).

Enfin, ces visions distinctes se traduisent par des conceptions divergentes du rôle des économistes. S'agit-il de « réparer » (*fix*) le monde et de se concentrer sur les détails pratiques de la mise en œuvre des politiques (DUFLO, 2017), à l'image d'un plombier ou d'un ingénieur qui répare des tuyaux fissurés ? Ou bien les économistes doivent-ils garder une distance critique par rapport au fonctionnement du système actuel, voire aller jusqu'à le remettre radicalement en question ?

Ces différentes positions épistémologiques (qui se présentent davantage comme un continuum plutôt qu'une opposition binaire) transparaissent dans les débats autour des RCT et se manifestent dans une série d'oppositions qui jalonnent les chapitres de ce livre : macro *versus* micro, biens publics *versus* biens privés, interventions sanitaires horizontales *versus* verticales, action publique *versus* marketing social, structure *versus* comportement individuel, attribution *versus* processus, etc.

L'épistémologie des RCT dans le domaine du développement

En théorie, les *randomistas* voient l'expérimentation précisément comme un antidote aux idées préconçues (RODRIK, 2009). Ce pragmatisme pourrait bien donner l'impression de rejeter le scientisme, mais le fait de revendiquer la supériorité d'une méthode reflète clairement une conception scientiste de la science (Picciotto, chap. 9, ce volume). Ce scientisme se manifeste de deux manières. Tout d'abord, les *randomistas* prétendent fournir des réponses universelles sur un grand nombre d'interventions de développement. En réponse à la question des particularités contextuelles, certains *randomistas*, comme Esther Duflo, soutiennent qu'il convient de considérer les RCT comme des « biens publics mondiaux » et de créer une instance internationale chargée de les multi-

plier (Savedoff *et al.*, 2006 ; Glennerster, 2012). Celle-ci constituerait ainsi une base de données universelle et jouerait le rôle de « chambre de compensation » apportant des réponses sur « ce qui fonctionne ou ne fonctionne pas » en matière de développement (Banerjee et He, 2008). Mais ce projet hégémonique (Bédécarrats *et al.*, 2019b) ne résout pas la question de l'hétérogénéité, qu'il s'agisse des pratiques ou des contextes d'intervention (voir notamment Spears, Ban et Cumming, chap. 6, ce volume).

Ensuite, ce scientisme se manifeste par un excès de confiance dans la technique, avec en quelque sorte un fétichisme du protocole théorique, censé garantir l'équilibre des échantillons et donc régler la question de l'attribution. La *mise en œuvre* du protocole sur le terrain est secondaire. Comme pour toutes les recherches – et plus encore pour les RCT, compte tenu des budgets en jeu, de la taille des échantillons, des contraintes de comparaison entre les groupes de contrôle et de traitement, et des risques de contamination –, la mise en œuvre des protocoles s'écarte nécessairement de ce qui est prévu en théorie et nécessite des ajustements, des aménagements et des compromis[8]. Dans de nombreux cas, la collecte de données des RCT va à l'encontre des hypothèses des théorèmes utilisés pour l'inférence statistique. Les ONG et les gouvernements qui travaillent dans le domaine du développement ne savent que trop bien que les interventions sur le terrain ne se déroulent jamais comme prévu (Mosse, 2004 ; Olivier de Sardan, 1995 ; 2021). Pourquoi les expérimentations devraient-elles être différentes ? Comme le montrent les différents chapitres de cet ouvrage, des écarts entre le protocole et la mise en œuvre peuvent être observés tout au long de la chaîne de production des connaissances.

– Des écarts au niveau de la construction des échantillons avec trois types de difficultés. La première difficulté réside dans les multiples biais entre les groupes de traitement et de contrôle (Ravallion, chap. 1, ce volume). Cela se traduit par une focalisation sur des populations très spécifiques, sans que cette spécificité soit mise en évidence par les *randomistas* (voir, par exemple, Wydick [2016] et Bédécarrats *et al.* [2019a] sur le microcrédit ; voir également Barrett et Carter [2014 : 75], Moatti, entretiens, ce volume). La deuxième difficulté réside dans le taux d'adhésion (*take up*) trop faible et, par conséquent, dans une différence insuffisante pour l'exposition à l'intervention. Le manque de puissance statistique affaiblit la capacité à tirer des conclusions, un problème qui, pour être résolu, nécessiterait des échantillons de taille peu réaliste et donc des budgets eux aussi irréalistes (McKenzie, 2012 ; Spears, Ban, et Cumming, chap. 6, ce volume). Un *take up* (taux d'adhésion) insuffisant peut également entraîner une transformation artificielle de l'intervention (voir le point suivant). Enfin, la « virginité » des zones de contrôle, condition souvent nécessaire à la comparaison, s'avère particulièrement complexe et pose des problèmes d'éthique et de faisabilité (Bédécarrats *et al.*, 2019b).

8. Notre expérience de réplication montre la difficulté qu'ont certains *randomistas* à reconnaître les difficultés pratiques que pose la réalisation d'une RCT idéale, dont l'équivalent n'existe pas dans la réalité (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

– Des écarts au niveau du type d'intervention, dont la mise en œuvre peut s'avérer très différente du « monde réel », comme le montrent Garchitorena *et al.* (chap. 5, ce volume) dans le domaine de la santé, ou qui risque même d'être transformée artificiellement pour favoriser un meilleur *take up* (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

– Des écarts au niveau de la collecte de données, puisque la priorité accordée aux considérations économétriques peut faire obstacle aux considérations statistiques. La statistique n'est pas seulement la science des chiffres, c'est avant tout une science de la collecte de données, dont la qualité est garantie par de multiples techniques (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

– Des écarts au niveau de l'interprétation des résultats qui, loin de se limiter à une comparaison de moyennes, comme le prétendent les *randomistas*, implique en réalité un éventail d'hypothèses implicites et tout un art de la rhétorique dont le pouvoir de persuasion est particulièrement manifeste (Labrousse, chap. 8, ce volume).

Au final, les contraintes de mise en œuvre de la méthode peuvent obliger les chercheurs à se concentrer sur des indicateurs intermédiaires, des périodes de courte durée, des populations ou des zones géographiques spécifiques et, ce faisant, à se limiter à un ensemble très restreint de questions ou à produire des résultats inutilisables (voir les chap. 5, 6 et 7 de ce volume sur différents secteurs avec de nombreux exemples ; voir aussi le cas de la santé publique [Moatti, entretiens, ce volume] et de la gouvernance [Natarajan, entretiens, ce volume]). L'importance disproportionnée accordée à la pureté théorique des protocoles et à la démonstration de la causalité au détriment de la *faisabilité* des protocoles et de la *qualité* des données est un point de friction majeur (bien que souvent implicite) dans les désaccords sur la hiérarchie des méthodes.

De notre point de vue, donner la priorité à la méthode plutôt qu'aux questions de recherche équivaut à « chercher ses clés perdues sous le lampadaire » parce qu'il n'y a que là que l'on voit quelque chose. D'une certaine manière, et pour paraphraser le titre d'un livre sur l'aide au développement (NAUDET, 1999), il s'agit de trouver des problèmes (les projets à évaluer) aux solutions (les RCT).

Les RCT et le « développement »

Comme le suggère Lant Pritchett (chap. 2, ce volume), le succès des RCT n'est que le symptôme d'une maladie plus grave : l'abandon, par une partie de la communauté de l'aide internationale, des politiques de développement transformatrices à grande échelle (nationale, régionale et même internationale), ainsi que des tentatives de transformer en profondeur les systèmes socio-économiques[9]. Il est donc utile d'examiner ces transformations dans le domaine de

9. Le jury du prix Nobel, dans son communiqué de presse, en témoigne : « Les lauréats de cette année ont introduit une nouvelle approche pour obtenir des réponses fiables sur les meilleurs moyens de lutter contre la pauvreté dans le monde. En résumé, elle consiste à diviser cette problématique en questions plus petites et plus faciles à traiter ».

l'aide pour mieux comprendre l'attrait des RCT et leur champ d'application. Le contraste entre la portée limitée des RCT et leur succès scientifique, médiatique et politique résulte à la fois d'une offre et d'une demande. Du côté de l'offre, nous avons déjà montré que les *randomistas* ont élaboré un *business model* scientifique inédit, dont le J-Pal est l'exemple le plus emblématique et le plus abouti, et qui associe des qualités qui se renforcent mutuellement : 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) (Bédécarrats *et al.*, 2019b). Aussi efficaces qu'elles soient, ces stratégies supposent néanmoins qu'il existe une *demande*. Si certaines méthodes, théories et technologies se révèlent efficaces, ce n'est pas en raison de leur supériorité scientifique, mais parce qu'elles parviennent à « galvaniser et rallier durablement des acteurs et des intérêts prêts à produire et à utiliser [les technologies en question] » (Callon, 2006a : 155).

Les RCT bénéficient ici d'un environnement qui leur est particulièrement favorable, et qu'elles entretiennent en retour. Elles n'auraient certainement pas eu le même succès à une autre époque. Le climat universitaire, notamment en économie, est propice à l'essor des RCT avec la défaite des écoles hétérodoxes centrées sur les structures sociales et les processus de domination, la recherche des fondements micros de la macro et le primat de la quantification et de l'économie dans les sciences sociales. La montée en puissance conjointe de l'économie comportementale et expérimentale, consacrée par l'attribution en 2002 du Nobel d'économie au psychologue Daniel Kahneman et à l'économiste Vernon Smith, puis à l'économiste Richard Thaler en 2017, illustre cette évolution de la discipline. Les RCT mobilisent très largement les préceptes de l'économie comportementale, et c'est d'ailleurs par leur intermédiaire que celle-ci s'est diffusée en économie du développement, jusqu'à y occuper aujourd'hui une place prépondérante (Fine *et al.*, 2016).

C'est également à la suite de transformations dans le domaine de l'aide que la demande de RCT est apparue. La fin de la guerre froide a favorisé une émancipation relative de l'aide publique au développement (APD) à l'égard du politique. Pendant cette période, la coopération technique et financière ne constituait souvent qu'un registre supplémentaire des rivalités entre blocs. Mais cette subordination de la coopération à la *realpolitik* a toutefois été battue en brèche après la chute du mur de Berlin. Dans le nouveau monde post-moderniste, les promoteurs de l'APD se sont retrouvés sous les feux de la rampe, sommés d'apporter la preuve de leur utilité dans un contexte de crise de l'aide, des Objectifs du millénaire pour le développement (OMD) et du *New Public Management* (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 grandes conférences internationales sur l'aide publique au développement, à Accra en 2008, puis à

Busan en 2011 et enfin à Addis Abeba en 2015. 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 (mobilisation compassionnelle et engagement moral ; Labrousse, chap. 8, ce volume). Leur (apparente) simplicité les rend aisément compréhensibles par les décideurs. Elles apparaissent donc comme un vecteur privilégié pour éclairer la décision publique. L'évaluation du programme *Progresa* au Mexique a constitué un prototype de cette méthode et un cas d'école de sa performativité[10] (Bédécarrats *et al.*, 2019b).

La crise de l'aide est également une crise de l'aide *publique* au développement. Alors que les efforts de financement de l'APD s'essoufflent, les investissements privés et les transferts de fonds internationaux prennent le relais (IFC, 2017). Les gouvernements ne sont plus qu'un organe parmi d'autres dans une « coalition d'acteurs » composée d'entreprises, d'ONG, et plus largement de la « société civile », de fondations et d'instituts de recherche. Renouant avec le philanthrocapitalisme de la période industrielle, les fondations jouent un rôle croissant, principalement dans le secteur de la santé, mais aussi dans l'innovation technologique, désormais présente dans la plupart, voire la totalité des secteurs du développement (voir aussi de Souza Leão et Eyal, 2020). Ces nouveaux acteurs et bailleurs de fonds modifient les *outils* de l'aide au développement. Non seulement le retrait de l'État en tant que responsable de la planification et du développement conduit à « penser petit » (Cohen et Easterly, 2010), mais, lorsqu'il va de pair avec la résurgence de la philanthropie, il ouvre la voie à un développement qui juxtapose la privatisation (des interventions et des acteurs), la marchandisation (des biens et des services fournis), mais aussi la compassion.

En faisant des pauvres des entrepreneurs aux pieds nus, le microcrédit, avec sa promesse d'un double résultat – réduction de la pauvreté et rentabilité ou au moins durabilité financière – a été un pionnier de la marchandisation. Cette marchandisation s'est ensuite développée sous le nom de *Bottom of the Pyramid* (BoP – bas de la pyramide en français), version *low-cost* de la théorie du *trickle down*, avec l'idée que la consommation des pauvres finira par constituer un moteur de croissance et de redistribution (Elyachar, 2012).

À cette motivation économique s'ajoute une « raison humanitaire » (Fassin, 2010). Face à des infrastructures publiques considérées comme moribondes,

10. Il est cependant édifiant de constater que ce programme a été un puissant outil de contrôle social et politique, rongé par le népotisme et la corruption (Crucifix et Morvant-Roux, 2018 ; Kidd, 2019). Par ailleurs, les lacunes de son évaluation expérimentale, notamment en termes de validité interne (Faulkner, 2014), ont précisément été les arguments utilisés par le nouveau gouvernement mexicain pour annoncer son retrait début 2019 (Encisco, 2019).

délabrées ou utopiques, et aux souffrances et besoins qu'elles engendrent, un devoir moral d'agir est en train d'émerger. Animés par un sentiment de compassion et d'urgence, les financiers et les praticiens – mais aussi les chercheurs – s'associent pour concevoir et tester un ensemble d'interventions à micro-échelle : ces « biens humanitaires », pour reprendre l'expression de REDFIELD (2012), tentent de répondre au mieux, de manière ponctuelle et temporaire, aux besoins considérés comme les plus urgents et les plus criants. Ces biens humanitaires visent à pallier les défaillances des gouvernements, et ils s'inscrivent dans cette double logique compassionnelle et économique, même si le volet économique n'exclut pas les mesures de redistribution (voir ci-dessous).

Dans cette nouvelle configuration, et même si le financement public des grandes infrastructures continue de représenter une part importante de l'aide internationale, les pouvoirs de décision et de planification des gouvernements cèdent progressivement la place aux fonds verticaux[11], aux fondations[12], aux entreprises privées[13] et aux nouveaux mécanismes financiers tels que les obligations à impact social. Les fondations, nouveaux acteurs en pleine croissance, sont vouées à jouer un rôle de plus en plus important (voir également Pritchett, chap. 2, ce volume). À l'instar de la fondation Ford, qui a soutenu l'essor des expériences aux États-Unis dans les années 1960, de nombreuses fondations jouent aujourd'hui un rôle moteur dans la généralisation des RCT dans le domaine du développement (à commencer par la mise en place de J-Pal ; JATTEAU, 2016 : 230). Le principe même des obligations à impact social, dont le remboursement aux investisseurs est conditionné par l'obtention de résultats sociaux précis, favorise une tendance similaire. Enfin, dans ce processus de privatisation du développement (privatisation des interventions et des acteurs), les ONG occupent une place de choix en tant que partenaires de mise en œuvre.

Loin des ambitions réformatrices et parfois idéalistes des générations précédentes d'acteurs du développement, les biens privés, marchands et humanitaires ont le mérite d'être réalistes et concrets, et d'offrir une solution pragmatique à des besoins considérés comme urgents. Leur déploiement n'est pas exempt de critiques – les plus connues étant sans doute les débats sur l'alimentation thérapeutique vue comme une pratique commerciale déloyale ayant un impact sur les systèmes agricoles locaux. Pourtant, du point de vue de leur objectif – résoudre un problème temporaire et individuel –, ils fonctionnent bel et bien (REDFIELD, 2012). Comme le soulignent plusieurs chapitres, et nous y reviendrons plus tard, ce sont précisément ces types de biens, en raison de leur ciblage individuel et de leur nature à court terme, qui se prêtent le mieux aux contraintes

11. Comme le Fonds mondial et Gavi, l'Alliance du vaccin (anciennement l'Alliance globale pour les vaccins et l'immunisation) dans le domaine de la santé.

12. La fondation Bill et Melinda Gates reste leader dans de nombreux sous-secteurs de la santé, mais aussi dans tout ce qui concerne les nouvelles technologies. Les fondations bancaires telles que Citi et Mastercard sont des acteurs de premier plan dans le domaine de l'inclusion financière.

13. Comme Nutriset pour les aliments thérapeutiques destinés à traiter la malnutrition et Vestergaard Frandsen pour les filtres à eau, les moustiquaires contre la mouche tsé-tsé et les moustiquaires imprégnées d'insecticides.

des essais randomisés. De même, les ONG restent les partenaires opérationnels de choix pour les *randomistas*, car elles sont plus flexibles, moins bureaucratiques, plus ouvertes à l'innovation et plus fiables que les gouvernements (COHEN et EASTERLY, 2010 ; WEBBER et PROUSE, 2018). Si les *randomistas* expriment le souhait de travailler davantage avec les gouvernements (BANERJEE, 2013), ils ont toutefois du mal à le faire (Pritchett, chap. 2, ce volume). En Inde, terrain d'étude privilégié pour les RCT, les témoignages apportés dans les entretiens de cet ouvrage par un haut fonctionnaire indien (Natarajan) et un ancien conseiller économique principal du gouvernement indien (Patnaik) suggèrent que l'impact des RCT sur l'élaboration des politiques est non seulement négligeable, mais aussi contre-productif, car les RCT détournent l'attention des problèmes réels et sapent le métier des économistes.

Cette transformation du domaine du développement étant intervenue bien avant les RCT, il serait exagéré de dire qu'elles en sont responsables (Morduch, chap. 3, ce volume), même si les effets d'éviction sont à prendre au sérieux (voir plus bas la section « La controverse peut-elle vraiment être évitée compte tenu des effets d'éviction ? »). Mais faut-il pour autant condamner sans réserve ces changements en dénonçant l'abandon de toute perspective réelle de réforme, la non-viabilité des interventions individuelles ponctuelles et l'illégitimité des acteurs privés qui ne sont pas démocratiquement responsables ? Ou faut-il apprendre à vivre avec eux de manière rationnelle, en considérant que, même si les tuyaux ont été mal conçus au départ ou sont sur le point de se casser – pour reprendre la métaphore du plombier/de l'ingénieur – cela vaut toujours la peine de réparer les fuites ? La réponse à cette question (rarement énoncée) éclaire de nombreux désaccords autour des RCT, ainsi que les différentes positions que l'on trouve dans cet ouvrage.

Éthique et RCT

La question éthique est récurrente avec les RCT, non seulement dans le domaine du développement, mais aussi en général (surtout en médecine). Si tout le monde s'accorde sur la nécessité d'aborder cette question de front, du moins sur le principe, ces mises en garde n'ont pas encore été suivies d'effet (dans ce volume, Abramowicz et Szafarz, chap. 10 ; et aussi Ravallion, chap. 1 ; Ogden, chap. 4 ; Bédécarrats, Guérin et Roubaud, chap. 7 ; Picciotto, chap. 9 ; Patnaik, entretiens, ce volume). Parmi les *randomistas*, cette prise de conscience reste marginale[14], comme si la foi dans les progrès scientifiques que les RCT peuvent apporter – et leurs retombées automatiques sur le plan des politiques et de l'amélioration du bien-être – suffisait à exempter les chercheurs de toute considération éthique.

14. Par exemple, parmi les 22 pages du chapitre « *Concerns about experiments* » de BANERJEE et DUFLO (2014), aucune n'aborde la question éthique, si ce n'est pour dire, en réponse au fait que la randomisation n'est pas une façon équitable de répartir le programme (car considérée comme un problème méthodologique, mais pas comme une question éthique), que « les responsables de la mise en œuvre pourraient trouver que la façon la plus simple pour […] présenter [le programme] à la communauté est de dire que [son] expansion […] est prévue pour les zones de contrôle de l'avenir » (*ibid.* : 101).

Alors que toute recherche soulève des questions éthiques, les RCT sont encore plus concernées que les études non expérimentales en raison de leur principe même (Teele, 2014), car elles présentent généralement une forme de manipulation de l'environnement de recherche (elles « tirent la queue du lion », pour reprendre l'expression utilisée par Deaton et Cartwright [2018 : 18]).

Les analyses critiques font également abstraction de ces considérations éthiques. Elles se contentent souvent de faire allusion au problème en ne donnant que bien peu de détails. Le chapitre d'Abramowicz et Szafarz (chap. 10, ce volume) fait figure d'exception en explorant en profondeur les implications du principe d'*équipoise*, c'est-à-dire l'exigence éthique selon laquelle une expérience impliquant des sujets humains doit révéler « un état de réelle incertitude de la part de l'investigateur clinique quant aux mérites thérapeutiques comparatifs de chaque bras de l'essai » (Freedman, 1987 : 141, cité par Abramowicz et Szafarz, chap. 10, ce volume). Les auteurs se demandent pourquoi les expérimentateurs en économie ignorent presque systématiquement ce principe, alors qu'il s'agit d'un pilier essentiel de la science médicale. Ils fournissent une série de pistes pour répondre à la question. Ravallion aborde également ce sujet (chap. 1, ce volume), en insistant sur l'importance de bien évaluer les risques et les informations déjà disponibles, et en montrant que le principe d'*équipoise* prend différentes formes selon les divers cas et types de randomisation (rationnement inévitable des traitements, randomisation conditionnelle, et essais d'équivalence). Il évoque également la proposition d'*expérience adaptative* avancée par Narita (2018) pour établir un optimum de Pareto entre les potentiels effets positifs et négatifs sur les participants en fonction des connaissances disponibles.

Cette quasi-négation des considérations éthiques par les *randomistas* est d'autant plus contestable qu'il existe différentes normes de bonnes pratiques, tant pour les RCT dans le domaine médical que pour la plupart des RCT en sciences sociales menées dans les pays du Nord. Les principes éthiques destinés à régir les essais randomisés impliquant des sujets humains ont été édictés dans des normes reconnues, en particulier la déclaration d'Helsinki en 1967 (WMA General Assembly, 2014), le rapport Belmont en 1974 et les lignes directrices internationales d'éthique du Conseil des organisations internationales et des sciences médicales (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979 ; Council for International Organizations and Medical Sciences, 2002). Ces normes établissent des principes clairs : le consentement éclairé, le principe « *do no harm* » (ne causer aucun préjudice), la fourniture d'une protection spécialement envisagée pour les populations vulnérables, l'analyse des risques et la surveillance réactive, pour n'en citer que quelques-uns.

Ces principes fondamentaux sont rarement respectés dans le domaine du développement (Abramowicz et Szafarz, chap. 10, ce volume). Comme Barrett et Carter (2014), nous détaillons quatre exemples de RCT qui illustrent les effets néfastes de cette négligence éthique. Dans le premier exemple, la RCT visait à démontrer les mécanismes de la corruption dans le cas de l'obtention du permis de conduire en Inde (Bertrand *et al.*, 2010). L'un des bras du traitement consistait à offrir une prime aux candidats qui obtiennent leur permis. Barrett et Carter

montrent que cette RCT a enfreint le code éthique du « *do no harm* » (ils parlent même de « conception irresponsable de la recherche ») de deux façons : non seulement le dispositif encourageait la corruption, mais il mettait également en danger la vie d'autrui en laissant le volant à des conducteurs potentiellement imprudents, puisque l'expérience a montré que le groupe traité prenait moins de leçons de conduite. Le second exemple concerne une RCT mise en place au Kenya pour tester « l'effet Rockefeller » (qui affirme que trop de ressources nuisent plus qu'elles ne sont bénéfiques) *via* un projet d'assistance à des groupes de femmes (Gugerty et Kremer, 2008). Les effets du projet se sont avérés négatifs (les femmes les plus pauvres ont été exclues des postes à responsabilité), confirmant ainsi l'hypothèse de Rockefeller. Cette RCT a causé un préjudice au sujet de l'expérience, un préjudice qui aurait pu être prévu, au moins comme étant une possibilité. Les femmes auraient donc dû être informées de cette éventualité afin qu'elles puissent décider de participer ou non à l'expérience (principe du consentement éclairé). Le troisième cas est une RCT impliquant des élèves de l'enseignement secondaire en République dominicaine. Il s'agissait de tester si des informations concernant les rendements de l'éducation sur le marché du travail, supérieurs à l'idée que s'en faisaient les élèves *a priori*, pouvaient les conduire à prolonger leurs études (Jensen, 2010). Ici, le problème éthique est le suivant : les informations fournies aux élèves du secondaire (estimées à partir de données non expérimentales) – qui sont à la fois surestimées (étant donné les biais d'endogénéité) et calculées en moyenne sans tenir compte des caractéristiques des élèves et des écoles – ont probablement conduit certains des élèves, sans doute les plus pauvres, à « surinvestir » dans l'éducation en se basant sur les bénéfices attendus. Sans compter l'effet de l'augmentation de l'offre des diplômés susceptible de peser sur les rendements futurs (effet d'équilibre général). Enfin, le traitement dans le cadre de la quatrième RCT consistait à accorder des crédits aux personnes rejetées par un prestataire de microcrédit en raison de la forte probabilité de défaut de paiement identifiée par son modèle de notation (Karlan et Zinman, 2009)[15]. Outre que cette stratégie faisait courir au groupe traité le risque d'être incapable de rembourser les prêts (avec les pénalités associées) et de se retrouver dans une situation potentielle de surendettement, le fait de ne pas l'en avoir informé est contraire au principe du consentement éclairé. Cette situation soulève un dilemme délicat puisque, si les participants avaient été informés du risque, leur comportement aurait probablement changé, compromettant ainsi la validité interne de la RCT. Aucune de ces failles n'a empêché ces quatre RCT d'être publiées dans des revues universitaires de premier plan, ce qui soulève également des questions de responsabilité des revues économiques dans le non-respect des normes éthiques (Abramowicz et Szafarz, chap. 10, ce volume).

D'autres exemples sont cités dans le présent ouvrage (Ogden, chap. 4 ; Abramowicz et Szafarz, chap. 10 ; Patnaik, entretiens, ce volume). Le recours

15. Une approche similaire consistant à inclure des sujets initialement jugés insolvables figure également dans Augsburg *et al.* (2015) ; elle est examinée par Bédécarrats, Guérin, et Roubaud, chap. 7, ce volume.

accru aux RCT, notamment par des institutions moins visibles et donc encore moins contrôlées sur le plan éthique, pourrait finir par ébranler les principes de base. Le cas d'une RCT en cours illustre ce point. Cette RCT, commandée par des bailleurs de fonds, a été mise en place pour tester la manière dont l'information affecte le comportement migratoire dans les zones rurales du Mali. Les participants ont visionné un court-métrage retenu parmi quatre films sélectionnés au hasard, illustrant différents scénarios de migration et de non-migration (vers l'Europe) : une migration réussie ; la souffrance, les mauvais traitements et en fin de compte une tentative de migration ratée ; une non-migration réussie ; et une comédie n'ayant rien à voir avec la migration à titre de placebo. Outre les problèmes liés à la communication aux participants des enjeux de ce test et à l'obtention de leur consentement éclairé, aucun des volets du test ne peut être considéré comme bénéfique pour les participants (violation du principe de *bienfaisance*). Les préférences des individus peuvent simplement être modifiées en fonction de ce que les commanditaires de la RCT considèrent comme étant bon pour eux (ou pour eux-mêmes). Le principe du « *do no harm* » n'est pas non plus respecté, car, après avoir visionné les films, certains participants pourraient décider d'émigrer au risque de mourir en Méditerranée ou d'être torturés dans les prisons libyennes. Enfin, la motivation politique du commanditaire semble évidente : freiner la migration africaine vers l'Europe. Il semble donc que cette RCT ait été conçue sans considération éthique sérieuse.

Au vu de la multitude d'exemples, il semblerait que la création de comités de protection des personnes dans de nombreuses institutions universitaires (*Institutional Review Boards*) n'ait rien fait pour remédier aux lacunes éthiques observées, ou du moins pas suffisamment (Barrett et Carter, 2020). On peut avancer deux raisons qui se renforcent mutuellement. La première est la difficulté de garantir simultanément la protection des sujets de l'expérience et la validité interne du protocole. La seconde est la compréhension imparfaite et le manque d'intérêt manifeste des *randomistas* pour le sujet. Face à ce que l'on pourrait appeler un dilemme éthique, ils penchent trop souvent du côté de l'impératif méthodologique. Pourtant, les garanties éthiques sont d'autant plus nécessaires dans les pays du Sud. Premièrement, le fait de ne pas informer les participants (principe du consentement éclairé), voire de mal informer délibérément les sujets humains pour assurer une stratégie d'identification théoriquement pure, est contraire au principe d'appropriation promu par les politiques de développement. Deuxièmement, les participants sont généralement des individus vulnérables, tant sur le plan économique (pauvres) que politique (sans voix), à qui il est plus facile d'imposer l'essai, quitte à les induire délibérément en erreur. Cette asymétrie est d'autant plus forte que les RCT sont de plus en plus souvent associées à des jeux de laboratoire grandeur nature, supervisés par de jeunes étudiants et des assistants de recherche rattachés à des universités de pays du Nord. Nous devons également nous pencher sur le choix de ces populations, notamment lorsque nous testons une hypothèse comportementale ou une théorie avancée par certains partisans des RCT (Banerjee et Duflo, 2011 ; voir Morduch, chap. 3, ce volume). Hormis l'argument selon lequel les pauvres des pays du Sud ont une rationalité spécifique, les arguments de

moindre coût et de moindre capacité à refuser de participer (problème récurrent avec les RCT menées dans les pays du Nord) en raison d'une méconnaissance de leurs droits et de rapports de force déséquilibrés (y compris vis-à-vis des expérimentateurs) semblent être des explications crédibles (Patnaik, entretiens, ce volume ; voir aussi TEELE, 2014), comme cela a déjà été observé lors de la « délocalisation » des essais cliniques médicaux (PETRYNA, 2007). Sans aller jusqu'à demander un « moratoire sur l'expérimentation » dans les pays du Sud (HOFFMANN, 2020), le sujet devrait au moins être traité en priorité.

Si les *randomistas* invoquent comme argument éthique l'amélioration à long terme du bien-être des populations grâce aux progrès scientifiques rendus possibles par les RCT, cette hypothèse est néanmoins loin d'être démontrée (Ravallion, chap. 1, ce volume). En définitive, outre la foi inébranlable dans la théorie de la technique au détriment de sa faisabilité (comme on l'a vu dans la section « L'épistémologie des RCT dans le domaine du développement »), il semble que trop souvent une hiérarchie de valeurs difficilement acceptable privilégie les résultats scientifiques plutôt que le bien-être des populations.

Quel est le champ d'application des RCT ?

Après avoir examiné de près les nombreuses limites des RCT, tant en termes de validité interne qu'externe, DEATON et CARTWRIGHT (2018) considèrent qu'elles restent néanmoins valables dans deux domaines : d'une part, tester une théorie ; d'autre part, évaluer ponctuellement et dans un contexte donné un projet ou une politique particulière, mais à condition que les problèmes potentiels de validité interne soient résolus et en tenant compte du fait que l'explication des résultats obtenus est souvent inappropriée. Les chapitres de ce livre confirment et développent cette analyse. Les évaluations randomisées ne sont possibles que pour un champ d'intervention très restreint, qui concerne le plus souvent des biens privés, marchands et humanitaires. Les RCT peuvent également être utilisées pour tester la théorie économique concernant les réactions comportementales aux interventions, et remettre ainsi en question certaines idées préconçues. Mais, en fin de compte, elles ne répondent ni à la question de l'*impact* tel qu'il est depuis longtemps défini dans le domaine de l'aide au développement, ni à la question de l'*explication* des effets mesurés.

Biens privés, marchands et humanitaires

Les conditions requises par les protocoles des méthodes randomisées les limitent à un spectre étroit que BERNARD *et al.* (2012) appellent les programmes « tunnels ». Ces derniers se caractérisent par des impacts à 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. Ils rejoignent les suggestions de Woolcock (2013) : les projets qui se prêtent à la randomisation doivent avoir une « *low causal density* », être peu contingents aux capacités de mise en œuvre (*implementation capability*) et comporter des résultats prédictibles.

Ce type de méthode n'est donc applicable qu'à des interventions simples ou ponctuelles, de courte durée, ciblant des individus. Concrètement, ces micro-interventions concernent essentiellement des biens et services privés, c'est-à-dire rivaux et exclusifs (voir Ravallion, chap. 1 ; Pritchett, chap. 2 et Picciotto, chap. 9, ce volume).

Dans le domaine de la santé, elles concernent les actions de prévention et de traitement de certaines maladies. Il peut également être question de filtres à eau, de moustiquaires, de formations et de systèmes de primes pour les professionnels de la santé, de consultations gratuites, de conseils médicaux par *Short Message Service* (SMS) et de micro-assurances. Cependant, les RCT ne répondent pas à la question de la gestion des systèmes de santé, qui sont par nature complexes et systémiques, et qui impliquent une main-d'œuvre qualifiée et motivée, une infrastructure, la fourniture de médicaments, etc. (Garchitorena *et al.*, chap. 5, ce volume). En matière d'assainissement, ces micro-interventions touchent à la distribution, la construction et l'utilisation de latrines. Là encore, les RCT ne répondent pas à la question de la gestion des flux de déchets humains : quel type de réseau d'assainissement ou de nettoyage utiliser, quel genre d'infrastructure, et quel type de régulation (Spears, Ban et Cumming, chap. 6, ce volume). En matière de réduction de la pauvreté, ces micro-interventions ont trait au microcrédit, à l'épargne, à la formation à l'entrepreneuriat et aux services d'éducation financière. Une fois de plus, les RCT ne répondent pas à la question des processus de création de richesses régionales ou sectorielles (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume), ni à la question plus large de l'accès aux services de base (Pritchett, chap. 2, ce volume). Dans le domaine de la gouvernance des administrations et des institutions publiques, ces micro-interventions prennent la forme d'inspections ponctuelles, de mesures d'incitation financière, d'audits par des tiers indépendants, de centres d'appels et de retours téléphoniques. Les RCT ne répondent pas à la question de la faible capacité de l'État, des bureaucraties centralisées souffrant d'un manque de confiance, des ressources limitées, des bureaucrates surchargés et des environnements de travail difficiles (Natarajan, entretiens, ce volume).

Contrairement à certaines analyses critiques (voir, par exemple, Berndt, 2015), les conclusions des RCT ne prônent pas nécessairement la marchandisation des biens privés (ce qui les rapproche davantage de la mouvance humanitaire). Dans le cas des moustiquaires imprégnées d'insecticides et des traitements vermifuges à forte élasticité de prix, les RCT ont précisément plaidé en faveur d'une distribution gratuite, considérée comme plus efficace que de les rendre payantes, ce qui remet en cause la croyance populaire du monde de la santé. Dans le cas du microcrédit, les RCT ont conclu que son impact sur la réduction de la pauvreté reste marginal et que la réduction de la pauvreté nécessite donc d'autres types d'interventions (Banerjee *et al.*, 2015c). Toujours dans le cas du microcrédit,

les RCT ont montré que les pauvres sont sensibles aux taux d'intérêt, en contredisant là aussi l'idée largement répandue selon laquelle l'accès est plus important que le coût, une croyance populaire parmi les organismes de microfinance et leurs financeurs, qui est invoquée pour légitimer des taux d'intérêt élevés (Morduch, chap. 3, ce volume).

Bien que ces résultats puissent être utiles, les sujets abordés restent limités par rapport à l'ensemble des questions de développement, de pauvreté et d'inégalités. Les conditions requises pour la mise en œuvre des RCT excluent donc un grand nombre de politiques de développement qui mettent en jeu des combinaisons de mécanismes socio-économiques et des boucles de rétroaction (effets d'émulation, d'apprentissage des bénéficiaires, d'amélioration de la qualité des programmes, effets d'équilibre général, etc.). C'est précisément le cas des biens publics (Ravallion, chap. 1, ce volume). Lorsque les interventions concernent des infrastructures et des systèmes de régulation, la manipulation expérimentale est impossible (Spears, Ban, and Cumming, chap. 6, ce volume).

Dans les termes de référence d'une étude commanditée sur le sujet, certains responsables du Department for International Development (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, PRITCHETT et SANDEFUR (2013b) aboutissent à des conclusions similaires[16]. Garchitorena *et al.* (chap. 5, ce volume) soulignent que 97 % des financements de la recherche en santé dans le monde sont consacrés au développement de nouvelles technologies (principalement pharmaceutiques), et que seuls les 3 % restants sont consacrés à la recherche sur la *mise en œuvre*, pourtant essentielle pour comprendre et améliorer les dysfonctionnements des systèmes de santé.

Évaluer l'impact ou tester le comportement ?

Comme le suggère Morduch (chap. 3, ce volume), les RCT visent en fait deux objectifs : mesurer l'impact et explorer « la nature des contrats économiques, des comportements et des institutions ». Il ajoute que ce deuxième type de « RCT exploratoire », moins controversé, est finalement le plus prometteur, puisqu'il représente un gain réel par rapport aux autres méthodes et donc un meilleur potentiel en termes de développement des connaissances.

Avec ce deuxième type de RCT, on ne cherche plus tant à mesurer l'impact des interventions caractéristiques de l'action publique ou de l'aide au développement, mais à expérimenter différents modes d'intervention et à mesurer des résultats en termes de *take up* de l'intervention. Selon Morduch (chap. 3, ce volume), ce

16. « Le champ d'application de l'approche de la "planification avec des preuves rigoureuses" en matière de développement est extrêmement limité » (PRITCHETT et SANDEFUR, 2013b : 1).

type de RCT est une source d'information, voire de « provocation », qui permet de remettre en question certains malentendus en matière d'économie du développement (comme la faible élasticité de la demande de microcrédit par rapport au prix, mentionnée plus haut) et de tester des innovations et la manière dont les personnes y réagissent. Par exemple, il peut permettre de tester différents calendriers de vente d'assurance pour la récolte afin de mieux comprendre les contraintes de temps et de liquidité ; ou de tester le rôle de l'information et de l'assistance dans l'utilisation des téléphones portables par les plus pauvres pour mieux connaître les mécanismes de partage au sein des ménages.

Si ces objectifs sont utiles et louables (à condition que les critères éthiques et de validité interne soient respectés et que les conclusions soient valables), on peut néanmoins se demander pourquoi les *randomistas* persistent à parler d'impact alors que de nombreuses RCT sont en fait de nature plus « exploratoire » et consistent à comparer différentes modalités d'une seule et même intervention, en se contentant souvent de mesurer les écarts de *take up*. Lorsqu'on analyse le secteur de l'assainissement, la conclusion est la même : les RCT semblent plus à même d'analyser les changements de comportement que de mesurer leur impact en tant que tel (Spears, Ban et Cumming, chap. 6, ce volume).

De fait, la question de l'impact reste souvent sans réponse. Depuis 1992, la plupart des acteurs du secteur de l'aide au développement s'appuient sur cinq critères définis par le Comité d'aide au développement de l'OCDE [Organisation de coopération et de développement économiques] (Development Assistance Committee, 2010), parmi lesquels figure celui de l'impact : « Effets à long terme, positifs et négatifs, primaires et secondaires, induits par une action de développement, directement ou non, intentionnellement ou non. » Les RCT ne peuvent cependant évaluer que l'*impact à court terme de chaînes causales courtes*. Il ne s'agit donc pas à proprement parler d'un impact tel que défini ci-dessus (voir également Picciotto, chap. 9, ce volume). Si l'on prend l'exemple des moustiquaires imprégnées d'insecticide, souvent considérées comme le fleuron des RCT (Ogden, chap. 4, ce volume), les RCT portent généralement sur le *take up* plutôt que sur l'impact, puisque les moustiquaires imprégnées d'insecticide sont considérées comme étant fondamentalement une « bonne chose ». Leurs effets à moyen et long terme sont pourtant controversés, puisque les moustiques ont fini par s'adapter génétiquement et que les systèmes de production locaux ont été détruits (Beisel, 2015). L'omission des effets à long terme et des effets collatéraux est tout aussi problématique dans le secteur du microcrédit (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

Ce type de RCT renvoie en fin de compte à la notion de « marketing social », terme très en vogue dans le monde du développement, qui rejoint tout à fait naturellement la circulation des biens privés, marchands et humanitaires, ainsi que la tendance comportementaliste décrites ci-dessus. Le marketing social consiste à appliquer des outils et des principes commerciaux à la conception, à la mise en œuvre et à l'évaluation de programmes de changement de comportement en vue d'obtenir des avantages individuels et de servir l'intérêt public (French *et al.*, 2010). Inspirées des sciences du comportement, les techniques

de marketing social comprennent les « *nudges*[17] », mais aussi des méthodes de marketing plus classiques (emballage, prix, identification des canaux et des lieux de distribution les plus appropriés, etc.) Le marketing social a vu le jour dans les années 1970 dans les secteurs sanitaire et social, notamment dans les pays du Sud. Il a aussi été introduit dans des domaines tels que la santé reproductive, la prévention du sida, la réhydratation en cas de diarrhées, l'assainissement, avant de s'étendre pour cibler les changements de comportement dans un grand nombre de secteurs (environnement, agriculture, éducation, gestion financière, consommation, etc.)

Mesurer *versus* expliquer

Si les RCT peuvent être capables de mesurer et de tester certains impacts et aspects des interventions, ils ne permettent toutefois pas d'analyser leurs *mécanismes*, ni leurs *processus* sous-jacents. Lorsque les causes de la pauvreté sont analysées sous l'angle des manques, comme c'est le cas dans les approches randomisées, la question des processus et des mécanismes n'est pas abordée (Shaffer, 2015). Pour pallier cette lacune de la théorie probabiliste de la causalité, il faudrait établir un « modèle causal » (Cartwright, 2010), une théorie cohérente du changement (Woolcock, 2013), une approche structurelle (Acemoglu, 2010) et une évaluation contextuelle de l'intervention (Ravallion, 2009a, et chap. 1, ce volume ; Pritchett et Sandefur, 2015).

Face à cette critique, les *randomistas* fondent désormais plus souvent leurs résultats sur des théories explicites du changement (Ogden, chap. 4, ce volume), basées en grande partie sur l'économie comportementale. Cette dernière permet de décoder toute la complexité des processus psychologiques et cognitifs, des luttes intérieures des individus, mais aussi de comprendre les multiples pratiques de « comptabilité mentale » (Thaler, 2015), ainsi que d'explorer et de tester la manière dont les comportements réagissent aux différentes interventions (voir la section « Évaluer l'impact ou tester le comportement ? »). Cependant, l'économie comportementale ne peut pas saisir la complexité des comportements atypiques, inattendus et « sous-optimaux », ce dernier terme impliquant qu'il faudrait définir ce qui relève de la norme. Il convient ici de distinguer deux niveaux : celui des comportements des individus, qui sortent parfois du cadre comportemental (Servet, 2018 ; Servet et Tinel, 2020), et celui des interventions, qui se déroulent rarement comme prévu.

En ce qui concerne le comportement des individus, ces derniers ne peuvent être réduits à de simples populations cibles, puisqu'ils sont des êtres sociaux et pluriels. Une personne ne se limite pas à sa capacité de refuser ou d'accepter, ni d'ailleurs à ses « biais » cognitifs ou sociaux. La rationalité et les motivations d'une population locale donnée sont le fruit d'une construction : elles découlent

17. En économie comportementale, un « *nudge* » (« coup de pouce » en français) désigne un dispositif de petite taille et peu coûteux qui vise à influencer le comportement des personnes de manière prévisible, sans qu'il ne s'agisse d'une obligation ni d'une interdiction formelle.

des normes et des réalités sociales et politiques et en sont le reflet. Elles renvoient à des formes préexistantes d'interdépendance, d'équilibre des pouvoirs et de structures sociopolitiques, mais aussi à des désirs et des aspirations. Si la dimension sociale est si imprévisible, elle ne doit pas pour autant être considérée comme un obstacle et une contrainte à éliminer à coups de « *nudges* ». Les populations locales ont parfois de bonnes raisons d'agir comme elles le font, surtout lorsque l'environnement global n'évolue pas. En effet, elles ont leurs propres conceptions et représentations du monde (et leurs propres théories du changement), ainsi que leurs propres connaissances et savoir-faire en matière de soins, de maladie et de bien-être, de propreté et de saleté, de finances, de pauvreté et de richesse, etc. Si certaines de ces représentations sont source de discrimination, il n'en demeure pas moins qu'elles façonnent les comportements. Elles reflètent par ailleurs des visions du monde particulières, qui ne sont pas nécessairement moins « optimales » que celles des chercheurs (voir l'exemple du microcrédit avec Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

Les interventions sont également complexes, puisqu'elles combinent plusieurs niveaux et acteurs. Les réalités locales façonnent, encadrent, contraignent et influencent les interventions (Mosse, 2004 ; Olivier de Sardan, 1995). C'est du moins le cas pour les trois secteurs représentés dans cet ouvrage. Dans le domaine de la santé mondiale, par exemple, « l'une des questions fondamentales […] consiste à savoir pourquoi les technologies connues – celles qui ont fait leurs preuves dans certains contextes – ne parviennent pas systématiquement à toucher les personnes auxquelles elles sont destinées » (Garchitorena *et al.*, chap. 5, ce volume). Pour y répondre, il faut forcément s'intéresser au fonctionnement des « systèmes » : que ce soit les systèmes de santé locaux, les systèmes d'organisation, les particularités des interactions entre les populations « cibles » et les prestataires de soins, etc. De même, l'assainissement et le microcrédit ne revêtent pas le même sens selon les personnes et englobent d'innombrables réalités, méthodes et formes de mise en œuvre. Cette diversité limite considérablement les possibilités de généralisation des RCT (Spears, Ban et Cumming, chap. 6 ; Bédécarrats, Roubaud et Guérin, chap. 7, ce volume). Ces trois secteurs ne sont sans doute pas les seuls à présenter une telle complexité et diversité.

Pourquoi une controverse scientifique est-elle nécessaire et pourquoi n'a-t-elle pas eu lieu ?

Comme nous l'avons vu, les RCT sont loin de faire l'unanimité et suscitent de nombreux débats et critiques. Alors qu'elles auraient dû déclencher une controverse scientifique, vitale pour le progrès scientifique et le débat démocratique, celle-ci n'a pas (encore) eu lieu. Comment l'expliquer ? Sans prétendre couvrir

le sujet de manière exhaustive, et compte tenu de son importance, il nous a semblé utile d'esquisser quelques pistes d'analyse empruntées au champ de l'épistémologie des sciences.

Contrairement à une vision naïve de la science, il faut garder à l'esprit que le progrès scientifique n'est pas toujours un processus rationnel et linéaire, où les méthodes et les résultats les plus efficaces et les plus utiles l'emportent systématiquement sur les autres, et où les connaissances attestées font l'objet d'un consensus. Les connaissances scientifiques sont également le fruit de l'histoire, de la société et de la politique, forgées par des succès et des échecs, des cycles, des débats et des désaccords, lesquels débouchent parfois sur des *controverses*, c'est-à-dire des divergences entre plusieurs parties exposées et débattues sur la scène publique.

Les controverses scientifiques ne doivent donc pas être vues d'un mauvais œil, comme étant la manifestation d'erreurs de raisonnement (où le « vrai » l'emportera finalement sur le « faux ») ou d'une ingérence indésirable de la politique ou d'intérêts autres que le progrès des connaissances (un domaine censé être exempt de toute subjectivité). Les controverses sont inhérentes à la production collective de connaissances, et permettent souvent l'émergence de progrès scientifiques majeurs. Tous les domaines scientifiques sont marqués par de grandes controverses parfois violentes (mémoire de l'eau, organismes génétiquement modifiés [OGM], « scandale de l'Eldorado », ondes gravitationnelles), mais parfois aussi étouffées dans l'œuf (Callon, 2006a).

Une controverse peut être définie comme un différend opposant deux camps et prenant à témoin un public composé de pairs scientifiques ou bien plus large (Lemieux, 2007). Si les divergences s'avèrent parfois virulentes, les participants sont néanmoins tenus de respecter les conventions du monde académique, notamment le principe d'égalité entre les participants, l'importance du raisonnement logique, la maîtrise de l'agressivité, et le respect du principe de la dignité des intervenants. Ces conventions restent cependant floues, et lorsqu'un adversaire est accusé d'abuser d'une position dominante ou de manquer de civisme, c'est souvent un moyen de renverser le rapport de force ou de disqualifier son rival.

Comme dans de nombreux domaines de la sociologie, la manière d'aborder les controverses diffère selon les écoles. Alors que certaines privilégient la logique et les preuves (Raynaud, 2018), d'autres se concentrent sur les croyances, les conventions sociales ou les rapports de force qui influent sur le contenu des arguments et l'arbitrage entre les logiques rivales (Akrich *et al.*, 2013). Mais, d'après les études scientifiques, quelle que soit l'école, le fait de contester les processus reflète une réalité sociale et historique, et met en lumière les rapports de force, les positions institutionnelles et les réseaux sociaux. Les controverses font avancer nos sociétés en modifiant les rapports de force, en redistribuant le prestige et les ressources, et en créant de nouvelles normes qui limiteront les actions et les positions futures (Lemieux, 2007).

Pour en revenir à notre question – pourquoi la controverse n'a-t-elle pas eu lieu ? –, les interprétations conceptuelles développées par Callon (2006a ;

2006b) fournissent quelques éléments de réponse. Premièrement, ce qui justifie ou non qu'une controverse ait lieu fait toujours l'objet d'accords négociés dans le cadre des processus de contestation. Dans notre cas, la communauté professionnelle des économistes du développement privilégie ce qu'elle estime relever de la recherche fondamentale, notamment la pureté statistique des essais randomisés et la maîtrise des biais d'identification des causes. Cet aspect prime ici sur des considérations, jugées secondaires par la communauté professionnelle, qui relèvent davantage du domaine appliqué et qui impliquent la reconnaissance des différentes « astuces du métier », des tactiques et des mises au point nécessaires à la mise en pratique de la méthode, mais aussi la reconnaissance du rôle des expérimentateurs et des sujets de l'essai, ainsi que des groupes de participants qu'il implique (Kabeer, 2019 ; Bédécarrats *et al.*, 2021). Cela nous ramène aux différences épistémologiques évoquées plus haut. La naissance d'une controverse implique donc la mise en place de forums suffisamment structurés pour que des discussions approfondies puissent avoir lieu. Sans ces espaces de débat, il est difficile de savoir qui parle et dans quel contexte, les mêmes acteurs pouvant défendre une argumentation partiale dans certains forums et, sans jamais se rétracter, faire des déclarations beaucoup plus équilibrées et prudentes dans les forums d'experts.

Éviter la controverse tout en écoutant et en s'adaptant

L'absence de dialogue public n'empêche pas les *randomistas* d'adapter leurs méthodes et pratiques (Ogden, chap. 4, ce volume), même si les réponses varient selon les groupes de chercheurs. Certains mettent leurs données à disposition, encourageant ainsi leur reproduction, pendant que d'autres reconnaissent le bien-fondé du pluralisme méthodologique et combinent les RCT avec d'autres méthodes. Alors que certains se concentrent en détail sur les mécanismes et les processus d'impact et utilisent des théories spécifiques (basées principalement sur l'économie comportementale), d'autres s'attaquent à la question de la validité externe et multiplient les études de cas dans différents contextes (le numéro spécial sur le microcrédit édité par Banerjee *et al.* [2015c] en est un exemple typique ; Bédécarrats, Guérin et Roubaud, chap. 7, ce volume), ou réanalysent *ex post* un certain nombre de RCT (Meager, 2019). D'autres encore s'intéressent sérieusement à la question du « penser petit » et se concentrent sur les programmes à grande échelle et les politiques nationales. Face à la faible influence sur les politiques publiques (Pritchett, chap. 2, ce volume), certains *randomistas* créent des organes dédiés, ou deviennent parfois eux-mêmes décideurs.

Reste à voir dans quelle mesure la *mise en œuvre* de cette nouvelle génération de RCT dans l'économie du développement peut faire face aux contingences du terrain et évaluer réellement des interventions plus complexes. Au risque de nous répéter, il convient de souligner cette obsession pour le protocole, jugé plus important que sa faisabilité et ses enjeux éthiques, et qui constitue l'un des points cruciaux du débat. Or, plus les programmes et les politiques

étudiés sont compliqués, plus il y a de risques que le protocole initial soit ajusté et que des compromis soient faits. Il ne s'agit pas seulement de revoir la technique, mais de se défaire d'une position épistémologique scientiste telle que définie ci-dessus.

La controverse peut-elle vraiment être évitée compte tenu des effets d'éviction ?

Si la controverse est indispensable, c'est aussi parce que la prétendue hiérarchie des méthodes produit des effets d'éviction, tant sur le plan des méthodes elles-mêmes (les autres étant discréditées) que sur celui du financement et des types d'interventions, avec par conséquent une dimension performative : le succès des RCT transforme le domaine du développement.

Pour illustrer cela, nous allons examiner deux exemples sur la question du financement. En Inde, une étude permettant d'évaluer véritablement l'impact de l'assainissement sur la mortalité infantile (l'indicateur le plus approprié, mais que les RCT ne permettent pas de mesurer statistiquement) coûterait environ 90 millions de dollars (sous certaines conditions ; Spears, Ban, et Cumming, chap. 6, ce volume). Une RCT classique coûte entre 500 000 et 1 500 000 dollars[18], et chaque RCT fait généralement l'objet d'une seule publication de recherche. Est-ce bien raisonnable, lorsqu'on sait que cette même somme permettrait de financer le système d'enquête statistique auprès des ménages d'un pays pauvre, et de donner lieu à une multitude d'études issues de ces données non expérimentales ? C'est l'une des questions cruciales que pose Patnaik (entretien, ce volume).

Le domaine de la santé illustre particulièrement bien les effets performatifs des RCT. Sans en être la raison première, les RCT ont contribué à l'essor des approches verticales dans le domaine de la santé (projets en silos). Axées sur le traitement individuel de maladies spécifiques, elles ont pris le pas sur les approches horizontales qui, elles, visent à développer des systèmes de santé complexes et intégrés (Garchitorena *et al.*, chap. 5, ce volume). D'autres études mettent en évidence les effets performatifs (et problématiques) du recours croissant aux RCT (Adams, 2016 ; Biehl *et al.*, 2014) : programmes non randomisables délaissés, modification des programmes pour les rendre plus facilement randomisables, priorité donnée à l'évaluation plutôt qu'à l'intervention elle-même (notamment en modifiant le travail du personnel de terrain : Adams, 2016). Les perturbations causées par les RCT et affectant la qualité des interventions ont été documentées dans d'autres domaines, tels que le microcrédit (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume) et la micro-assurance (Quentin et Guérin, 2013).

18. Il n'existe à notre connaissance aucune estimation précise du coût des RCT, mais Pamiès-Sumner (2015) donne quelques approximations. Voir également Ravallion, chap. 1, ce volume.

Quelles sont les alternatives de recherche ?

Notre objectif n'est pas de rejeter les RCT, car elles constituent une méthode prometteuse… parmi d'autres. Cependant, elles doivent être menées dans les règles de l'art, prendre au sérieux la question de la faisabilité et des enjeux éthiques en s'alignant sur les bonnes pratiques du monde médical, et se conjuguer à d'autres méthodes. Si les RCT conviennent à certaines politiques précisément définies, d'autres méthodes peuvent et doivent être utilisées, comme le montrent plusieurs chapitres de ce livre. Pour les projets que les RCT peuvent (en partie) couvrir, il convient de combiner les méthodes.

Une alternative à l'étalon-or consiste à adopter une approche pragmatique en définissant les questions de recherche et les outils méthodologiques requis au cas par cas en fonction des connaissances préalables disponibles, de la conception de l'intervention et des particularités des contextes, en liaison avec les différentes parties prenantes, qu'il s'agisse des opérateurs de terrain, des bailleurs de fonds, des gouvernements ou des populations locales largement négligées.

Ces méthodes alternatives s'appuient également sur une série de méthodes fondées sur l'interdisciplinarité et la reconnaissance des différentes manières de produire des preuves, tant quantitatives que qualitatives. Ces approches ne visent pas à fixer des lois universelles, mais à expliquer les liens de causalité spécifiques à un moment et à un lieu particuliers. Si le recours aux méthodes mixtes est souvent préconisé, que ce soit par les chercheurs[19] ou les institutions (voir Rioux, entretiens, ce volume)[20], on remarque qu'il y a un décalage avec la pratique, où cela se fait finalement peu. Alors que certains *randomistas* reconnaissent publiquement la légitimité des méthodes alternatives (Ogden, chap. 4, ce volume), ils ne tiennent souvent pas compte des résultats des méthodes non randomisées, ce qui semble contredire leur apparente ouverture d'esprit (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume).

Dans le domaine de la santé mondiale, les interventions sont d'une telle complexité que la randomisation est souvent impossible et laisse donc place à des méthodes non expérimentales et quasi expérimentales, plus appropriées. Comme le montrent Garchitorena *et al.* (chap. 5, ce volume), il existe de nombreux exemples de méthodes alternatives et complémentaires aux RCT, même si ces dernières restent utiles pour certaines interventions spécifiques. Ces méthodes alternatives présentent quelques particularités : elles reposent sur la théorie de la complexité (avec un système de santé global, plutôt que des

19. Voir, par exemple, les deux ouvrages mentionnés dans l'introduction des éditeurs (Cohen et Easterly, 2010 ; Teele, 2014), dont la plupart des chapitres et des déclarations liminaires insistent sur la nécessité de recourir à des méthodes mixtes. Voir également Camfield et Duvendack (2014).

20. Voir Picciotto, chap. 9, ce volume, à propos du monde de l'évaluation dans le développement. Voir aussi Pamiès-Sumner (2015) pour l'AFD, et les travaux du Centre of Excellence for Development Impact and Learning (CEDIL) (White et Masset, 2018) pour le DFID.

composantes fragmentées), combinent des méthodes et des échelles d'analyse, puisent autant que possible dans les systèmes statistiques nationaux et s'intéressent non seulement à l'impact, mais aussi à l'efficacité (en introduisant dans l'analyse des réalisations [*outputs*] et des résultats [*outcomes*], mais aussi des moyens [*inputs*] et des processus).

Outre les exemples cités dans le livre, nous soulignons également la nécessité de mener des méta-analyses et des réplications. Elles commencent à faire leur apparition en économie du développement, mais sont encore trop rares (Camfield et Duvendack, 2014). Ces réplications peuvent également être qualitatives et consistent à revisiter un terrain, comme cela a été fait au Maroc et au Bangladesh (Kabeer, 2019 ; Morvant-Roux *et al.*, 2014). Les méthodes qualitatives (entretiens semi-structurés, groupes de discussion, observation participante, ethnographie, études de cas, monographies, etc.) peuvent servir plusieurs objectifs : contextualiser les interventions, élaborer des hypothèses originales, identifier des phénomènes nouveaux et inattendus et analyser les interventions dans leur ensemble, en étudiant la complexité des liens de causalité et les nombreuses interactions dynamiques et contradictoires entre différentes entités de manière spécifique au lieu. Lorsque l'on est confronté à des chaînes causales complexes, ce qui est le cas de nombreuses interventions, les méthodes qualitatives sont souvent la seule façon de traiter réellement l'épineuse question de la causalité (White et Masset, 2018). Pourtant, elles sont souvent critiquées (à tort) pour leur incapacité à « démontrer » les résultats, et sont aussi utilisées de manière superficielle et non rigoureuse. Labrousse (chap. 8, ce volume) illustre cette dérive en évoquant le *storytelling* – un type de récit destiné à étayer un argument, mais qui n'a aucun pouvoir de démonstration – que certains *randomistas* utilisent à mauvais escient pour interpréter des résultats quantitatifs. La seule norme valable est finalement celle qui fait « bon usage des bonnes preuves » (Spears, Ban, and Cumming, chap. 6, ce volume).

Pour résumer et conclure cette introduction, nous pensons que certains principes clés devraient guider la recherche pour le développement, non pas en tant qu'alternatives aux RCT, mais en leur accordant une place à proportion congruente. Ces principes ne sont probablement pas révolutionnaires… mais leur mise en œuvre serait déjà un grand pas pour les sciences humaines et sociales. Premièrement, et pour passer du général au spécifique, la recherche doit être guidée par les grandes questions à traiter plutôt que par les méthodes pour lesquelles il faut trouver des applications. Pour paraphraser une citation célèbre : « Ne vous demandez pas ce que vous pouvez faire avec une RCT, mais demandez-vous plutôt ce qu'une RCT peut faire pour votre recherche ! » Deuxièmement, nous devons dépasser l'obsession de l'impact causal[21], qui domine la communauté des économistes du développement depuis la fameuse révolution de la crédibilité (Angrist et Pischke, 2010). Les analyses de données non expérimentales, les descriptions fouillées (*thick description*), les récits analytiques (*analytical narra-*

21. Ruhm (2019) qualifie cette obsession de « police de l'identification », qu'il suggère de « mettre aux fers ».

tives) sont autant d'autres questions et d'approches de recherche au moins aussi importantes pour faire progresser les connaissances, surtout si l'on considère que la pauvreté n'est pas seulement un problème de privation, mais aussi et surtout le résultat de rapports sociaux et de pouvoir. Troisièmement, en ce qui concerne les approches quantitatives, il est indispensable de rééquilibrer les efforts de recherche pour englober d'autres composantes de la chaîne analytique : les gains éventuels en termes d'attribution causale (en théorie, puisque différents chapitres de ce livre montrent que rien n'est garanti en pratique dans ce domaine) et le surinvestissement dans ce domaine ont relégué au second plan d'autres aspects tout aussi importants. Il y a tout d'abord la question de la qualité des données, trop souvent sacrifiée par manque d'intérêt et de compétence, ce que certains des plus grands *randomistas* commencent à reconnaître (Dillon *et al.*, 2020). Ensuite, il y a la multiplication des réplications qui s'attaquent de front au diagnostic méticuleux des données, et à leur inclusion dans les critères d'évaluation des revues académiques des pairs. Dans le même temps, il convient d'accorder une plus grande attention à la question des plans d'échantillonnage, car, s'ils sont complexes, leurs conséquences sont trop souvent négligées. Ces négligences entraînent la sous-estimation de la variance des estimateurs et la prise en compte d'impacts jugés statistiquement significatifs, alors qu'ils ne le sont pas (Gibson, 2019) pendant que d'autres le sont, mais que les RCT, pas toujours performantes, ne parviennent pas à identifier. Quatrièmement, il est temps de mettre réellement en pratique deux recommandations approuvées par tous, mais qui demeurent pour l'instant des paroles en l'air sans aucun résultat tangible : une prise en compte réelle des questions éthiques et la combinaison de méthodes qualitatives et quantitatives[22]. Le plaidoyer pour les méthodes mixtes et autres approches (MMA) est un vrai sport de combat (*Mixed Martial Sports*[23]) ! Enfin et surtout, il est temps de reconnaître une fois pour toutes que les expérimentations randomisées ne sont pas l'étalon-or de l'évaluation. L'orgueil démesuré qu'affiche une partie du mouvement pro-RCT pousse la recherche dans une impasse. C'est pourquoi modérer cette démesure s'avère essentiel et ne peut qu'être bénéfique. Il est également crucial de tirer les leçons du passé et de tenir compte des recherches antérieures, au moins à deux égards : les faiblesses des RCT (Heckman, chap. 12, ce volume) et les résultats des méthodes autres que les RCT. Autrement, toutes les tentatives évoquées jusqu'ici visant à écouter les critiques et à s'adapter sont vouées à l'échec, ou, pour paraphraser la célèbre expression de Lampedusa (1960) dans son roman, *Le Guépard*, « pour que tout reste comme avant, il faut que tout change ».

La consécration du prix Nobel d'économie conduira-t-elle les *randomistas* à mieux apprécier les avantages des différentes méthodes ou, au contraire, vont-ils

22. Comme le soulignent van der Meulen Rodgers *et al.* (2020) dans leur éditorial du numéro spécial du *World Development* sur les RCT, les contributeurs sont très nombreux à réclamer une triangulation, un pluralisme et une collaboration (tant au sein de la communauté scientifique qu'entre les universités, les bailleurs de fonds et la société civile).

23. Il s'agit d'un clin d'œil au documentaire réalisé par Pierre Carles en 2001 sur les travaux du sociologue Pierre Bourdieu, intitulé *La sociologie est un sport de combat*.

en profiter pour consolider leur position déjà quasi hégémonique ? Seul l'avenir le dira[24], mais nous tenons à souligner que la fin de l'étalon-or et de la quête de l'« incontestable », caractéristiques de la supériorité revendiquée par les *randomistas*, appelle une rupture épistémologique, mais aussi l'avènement de cette controverse que nous souhaitons tant. En se basant sur l'examen des controverses autour du changement climatique, LATOUR (2012) préconise de construire des espaces de débat ainsi que des méthodes pour discuter et débattre des différentes formes de connaissances scientifiques (dans toute leur diversité), et non scientifiques, en veillant à ce que les fondements idéologiques et politiques de ces multiples formes de savoirs ne soient ni rejetés ni occultés, mais explicités et débattus (EGIL, 2015). Nous pensons que ce projet, aussi ambitieux soit-il, est absolument nécessaire sur les plans scientifique et démocratique, si l'on veut véritablement améliorer les politiques de développement.

D'ailleurs depuis le début de l'année 2020, la pandémie de Covid-19, dont le monde n'est toujours pas sorti deux ans plus tard à l'heure d'écrire ces lignes, est venue éclairer d'un regard neuf le débat global autour des RCT, apportant son lot d'éléments nouveaux pour alimenter la plus que jamais nécessaire controverse scientifique sur le sujet. Les malheurs du monde offrent une extraordinaire « expérience naturelle » pour en apprécier la véritable contribution. Alors que la pandémie constitue le choc de pauvreté le plus brutal jamais connu à l'échelle mondiale en temps de paix, et que les RCT ont été primées pour leur contribution majeure à sa réduction, quel a été leur apport pour lutter contre cette catastrophe ? De plus, la pandémie de Covid-19, dans sa double dimension sanitaire et économique, ouvre une voie royale pour apprécier les mérites respectifs des RCT dans le domaine du développement et des essais cliniques (par exemple sur les vaccins). Par ailleurs et loin des fureurs du monde, l'attribution en octobre 2021 du prix Nobel d'économie à Joshua Angrist, Guido Imbens et David Card semble apporter de l'eau au moulin à l'obsession expérimentale, même si leur approche de l'inférence causale apparaît moins étriquée que celle des *randomistas*, en redonnant leur lettre de noblesse aux « expériences naturelles ». Il n'est pas question de tirer les enseignements de ce méga-épisode, en laissant la question ouverte, offerte à la sagacité du lecteur, et peut-être l'objet d'un second tome à venir, mais simplement d'annoncer qu'il ne fait que conforter les conclusions de cet ouvrage, tout en l'éclairant d'un jour nouveau.

24. Si l'on se base sur notre propre expérience, nous avons bien des raisons d'être pessimistes. En effet, il est de plus en plus difficile de publier des articles critiques dans les grandes revues universitaires et de trouver des interlocuteurs pour débattre de la contribution effective des RCT dans le domaine des politiques politiques. Si certains sont tentés d'émettre des opinions critiques, ou nuancées, vis-à-vis de la puissante et fameuse nouvelle doxa, ils se gardent bien de le faire, par autocensure ou pour des raisons institutionnelles.

Grandes lignes de l'ouvrage

Le livre est structuré comme suit. La première série de chapitres offre un aperçu des RCT dans le domaine du développement (à quels types de questions permettent-elles de répondre ou pas ?) ainsi que divers points de vue sur le potentiel de cette méthode (**Partie 1. Que peuvent les RCT ?** : chap. 1-4). La deuxième série de chapitres (**Partie 2. Perspectives sectorielles**) porte sur les analyses sectorielles dans le domaine de la santé (chap. 5), de l'assainissement (chap. 6) et du microcrédit (chap. 7), et pose les questions suivantes : quels enseignements avons-nous tiré des RCT dans chaque secteur et quelle est la contribution des autres méthodes ? La troisième série de chapitres (**Partie 3. Économies politiques**) propose des pistes de réflexion en matière d'économie politique, à la fois en ce qui concerne spécifiquement la rhétorique des *randomistas* (chap. 8) et plus généralement en inscrivant les RCT dans le champ et l'histoire de l'évaluation des politiques de développement (chap. 9). La quatrième et dernière série de chapitres (**Partie 4. [De quelques] Pistes de réflexions [ciblées] : éthique et méthode**) approfondit les propositions d'amélioration évoquées dans les chapitres consacrés aux secteurs, en mettant l'accent sur des aspects spécifiques, à commencer par la question de l'éthique, dont nous avons déjà souligné l'importance et la nécessité (chap. 10), puis en explorant les améliorations statistiques, l'utilisation des *a priori* (*priors* ; chap. 11) et la non-adhésion comme source d'information (chap. 12) ; sachant que cette partie fournit quelques éléments pour mieux faire dans le champ des RCT, et non un véritable dépassement, dont de nombreuses pistes sont explorées au fil des chapitres, y compris dans cette introduction. En guise d'épilogue, James Heckman propose une nouvelle relecture de son article phare de 1992, à la lumière de la nouvelle vague des RCT dans le domaine du développement. Il démontre que la plupart de ses conclusions d'alors, qui portaient sur la première génération des expérimentations randomisées dans le domaine de la politique sociale aux États-Unis (le « Premier Réveil » selon ses termes), sont toujours valables. Il appelle la nouvelle génération d'économistes à réfléchir et à tirer les leçons du passé. Enfin, l'ouvrage se conclut par une série de trois entretiens avec des décideurs politiques de premier rang, en laissant le point de vue de la recherche pour adopter une perspective de politique publique. Ces entretiens posent la question de l'utilisation, de l'utilité et des réponses apportées par les RCT pour prendre des décisions dans le monde réel. Le premier est un entretien croisé avec les présidents-directeurs généraux de nos institutions respectives, spécialisées dans le domaine du développement : l'aide pour l'Agence française de développement (AFD) (Rioux) et la recherche pour l'Institut de recherche pour le développement (IRD) (Moatti) ; il propose la vision d'un pays du Nord (la France). Les deux autres entretiens font intervenir des décideurs, confrontés quotidiennement à l'élaboration et au suivi des politiques économiques en Inde, terrain d'application privilégié des RCT dans les pays du Sud (Natarajan, haut fonctionnaire indien ; et Patnaik, ancien conseiller économique principal du gouvernement indien). En amont, Angus Deaton revisite, dans son introduction,

avec ses « onze variations », ses propres réflexions, à la lumière des contributions de cet ouvrage.

Remerciements

Cette introduction, dont les opinions sont de la seule responsabilité des auteurs, a été largement nourrie par les discussions lors de l'atelier qui a eu lieu à Paris le 17 mars 2019 et auquel ont participé la plupart des auteurs, ainsi que par les commentaires reçus d'Agnès Labrousse et d'Ariane Szafarz, que nous remercions chaleureusement.

Prologue
La randomisation sous les tropiques revisitée
Un thème et onze variations

Angus DEATON

Si les économistes du développement ont recours aux évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) depuis une vingtaine d'années[1], les économistes spécialisés dans les politiques de protection sociale aux États-Unis les emploient, eux, depuis bien plus longtemps. Grâce à ces années d'expérience, les discussions ont gagné en richesse et en nuances. Les partisans comme les détracteurs ont appris les uns des autres, du moins dans une certaine mesure. Comme à leur habitude, les chercheurs semblent réticents à tirer les leçons des erreurs commises par d'autres ; les enseignements de la première vague d'expérimentations, dont beaucoup ont été présentés par James Heckman et ses collaborateurs[2] il y a vingt-cinq ans déjà, ont souvent été négligés lors de la deuxième vague. Dans ce prologue, je ne cherche pas à reprendre l'ensemble des questions que j'ai déjà abordées ailleurs (DEATON, 2010a ; 2010b ; DEATON et CARTWRIGHT, 2018), ni à résumer le débat qui anime la sphère économique depuis longtemps. Je me concentre plutôt sur quelques-unes des questions qui occupent une place

1. L'attribution du prix Nobel à Abhijit Banerjee, Esther Duflo et Michael Kremer a été annoncée au moment de la révision de ce prologue. Le prix permettra, comme c'est déjà le cas actuellement, de donner davantage de visibilité au débat sur les avantages et les inconvénients de la conduite des RCT axées sur le développement économique. La presse s'est largement fait l'écho de préoccupations de fond, notamment sur le plan éthique. Ce débat médiatique révèle également de fausses idées très répandues, tant chez les détracteurs que chez les défenseurs, sur le fonctionnement réel des RCT, notamment des idées selon lesquelles les RCT garantissent que les groupes de traitement et de contrôle sont similaires avant le traitement, et qu'elles peuvent démontrer un lien de causalité.

2. Heckman (chap. 12, ce volume), qui est une version mise à jour de HECKMAN (1992), et HECKMAN et SMITH (1995). Voir également MANSKI et GARFINKEL (1992), qui contient la version de 1992 du document de Heckman, une excellente introduction générale de Manski et Garfinkel, et plusieurs autres documents toujours pertinents.

importante dans cet ouvrage critique et qui méritent, me semble-t-il, d'être réexaminées.

Je ne remets pas en cause l'utilité des RCT, mais, à mon sens, c'est une erreur de faire passer la méthode avant le fond. J'ai écrit des articles à partir de RCT (Deaton, 2012 ; Deaton et Stone, 2016). Comme d'autres méthodes de recherche, elles sont souvent utiles, mais présentent également des dangers et des inconvénients. Tout préjugé méthodologique nous lie inévitablement les mains. Le contexte revêt toujours une grande importance et il nous faut adapter nos méthodes au problème posé. Il est inexact d'affirmer qu'une RCT, lorsqu'elle est réalisable, sera toujours meilleure qu'une étude non expérimentale (*observational study*). Cela ne devrait pas prêter à controverse, mais, si j'en juge d'après la rhétorique employée dans la littérature, les propositions suivantes sont encore mal comprises, en particulier la seconde : (a) les économistes rencontrent avec les RCT les mêmes problèmes d'inférence et d'estimation qu'avec d'autres méthodes, ainsi que d'autres problèmes qui leur sont propres, et (b) aucune RCT ne peut légitimement prétendre avoir établi un lien de causalité.

Mon propos est le suivant : les RCT ne bénéficient pas d'un statut spécial, elles n'échappent pas aux problèmes d'inférence auxquels les économètres ont toujours été confrontés. Il n'y a rien qui ne puisse être accompli uniquement par les expérimentations randomisées. Tout comme leurs avantages ne sont pas l'apanage unique des RCT, leurs faiblesses non plus, et je m'efforcerai de le souligner. Il n'y a pas d'étalon-or. Il y a de bonnes et de mauvaises études, ni plus ni moins. Ce sur quoi je veux surtout insister, ce sont les dangers éthiques liés à la réalisation de RCT dans les pays pauvres, mais je garde ces remarques pour la fin.

Les RCT sont-elles le meilleur moyen pour apprendre et accumuler des connaissances utiles ?

Parfois oui, parfois non. Prétendre qu'une méthode est la meilleure, sous prétexte qu'elle est réalisable, ne rime à rien. Je n'ai jamais compris pourquoi le Abdul Latif Jameel Poverty Action Lab (J-PAL) se cantonnait aux RCT, se laissant ainsi accuser d'être plus (ou aussi) intéressé à faire du prosélytisme pour les RCT qu'à réduire la pauvreté. Mais, comme le fait remarquer Ogden (chap. 4, ce volume), les membres du J-PAL ont recours à un large éventail de techniques ; le J-PAL n'est peut-être donc que la branche RCT d'une entreprise plus vaste. Ravallion (chap. 1, ce volume) a parfaitement raison lorsqu'il affirme que la meilleure méthode est *toujours* celle qui fournit les réponses les plus convaincantes et les plus pertinentes au regard du contexte en question. Nous avons tous nos méthodes préférées, qui, selon nous, sont trop peu utilisées. Pour ma part, ce sont les tableaux croisés et les graphiques qui collent aux données. Le

plus dur est de décider ce qu'il faut y mettre et comment traiter les données afin d'apprendre quelque chose que nous ne savions pas auparavant, ou qui nous fait changer d'avis. Une image ou un tableau croisé bien construit peut saper la crédibilité d'une histoire de causalité largement répandue, ou, au contraire, renforcer la crédibilité d'une nouvelle histoire. De telles preuves en disent plus sur les causes qu'un article, dont le titre comporte le mot « causalité ». L'art consiste à savoir ce qu'il faut montrer. Mais je ne cherche pas à ce que tout le monde travaille aussi de cette manière.

Imposer une hiérarchie des preuves est à la fois dangereux et non scientifique. *Dangereux*, car des résultats dignes d'être considérés, et qui pourraient être cruciaux, sont automatiquement écartés. Les résultats d'une RCT sont pris en compte même si la population qu'elle couvre est très différente de celle où elle sera utilisée, si elle ne repose que sur quelques observations, si de nombreux sujets ont abandonné ou refusé d'y participer, ou s'il n'y a pas de mise en aveugle et que l'on peut s'attendre à ce que le fait de savoir que l'on participe à l'expérience en modifie les résultats. Il semble logique d'écarter des essais randomisés en raison de ces défauts, mais cela ne suffit pas si cela conduit aussi à exclure des résultats non randomisés qui seraient plus instructifs. Selon la hiérarchie, un résultat non randomisé n'est pas un résultat du tout, ou du moins pas un résultat « rigoureux ». Une étude non expérimentale sera rejetée même si elle est bien conçue, et ne présente pas de source évidente de biais en faisant appel à un très large échantillon de personnes pertinentes.

Non scientifique, car la communauté professionnelle se voit collectivement exonérée de l'obligation de concilier les résultats des différentes études ; les études non expérimentales sont considérées comme erronées simplement parce qu'elles ne sont pas randomisées. Une telle négligence dans l'utilisation des connaissances est relativement rare en économie, même s'il arrive bien couramment que les auteurs d'études randomisées ne citent pas les travaux non RCT, qu'ils les rejettent comme étant « anecdotiques » ou qu'ils soient incapables de dissocier la corrélation de la causalité (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume), mais il existe des exemples bien pires dans d'autres domaines, tels que la médecine ou l'éducation. Or, les économistes accordent souvent un poids particulier aux résultats issus de RCT sur la seule base de la méthodologie ; ces études sont considérées comme « crédibles », alors qu'elles ne font ni référence aux contextes, ni ne tiennent compte des alternatives.

L'économie est un domaine ouvert dans le sens où les études de qualité qui produisent des résultats nouveaux, importants et convaincants sont généralement jugées sur leurs mérites. Il convient toutefois de veiller à ce que le mérite ne serve pas de façade aux préjugés méthodologiques. Lorsque j'entends certains arguer que les RCT ont fait leurs preuves en débouchant sur de bonnes études, je veux être rassuré sur le fait que l'utilisation de la randomisation ne constitue pas en soi une mesure de la valeur et que l'argument n'est pas tout simplement circulaire.

L'inférence statistique est plus simple avec les RCT qu'avec d'autres méthodes

Ce malentendu est à l'origine de nombreuses incompréhensions. Il existe un problème qui attire rarement l'attention : les RCT, encore plus que les études non expérimentales, impliquent souvent des auteurs qui collectent des données, qui suivent les répondants dans le temps et qui identifient et traitent les cas aberrants. Ces tâches sont loin d'être simples et demandent beaucoup de temps ainsi que des compétences spécialisées que tous les économistes ne possèdent pas. Il est probable que les problèmes de collecte ou de traitement des données soient bien plus répandus que les erreurs d'inférence statistique (Bédécarrats, Guérin et Roubaud, chap. 7, ce volume), car ces questions sont loin d'être simples.

En matière d'inférence, l'argument de la simplicité comporte deux éléments. Premièrement, la randomisation garantit que les deux groupes, de traitement et de contrôle, sont globalement identiques avant le traitement, de sorte que toute différence observée entre eux après le traitement résulte nécessairement de ce dernier. Deuxièmement, l'inférence statistique nécessite le calcul d'une valeur *p* pour déterminer la différence entre deux moyennes, une procédure simple qui est enseignée dans les cours de statistiques de base. Ces deux éléments de l'argumentation sont faux.

Très tôt, Ronald Aylmer Fisher a compris que la randomisation ne permet pas d'équilibrer les observations entre les traitements et les contrôles, ce que ne tardent pas à constater tous ceux qui mènent des RCT. Ravallion (chap. 1, ce volume), qui a longtemps suivi les RCT à la Banque mondiale et ailleurs, affirme que ce malentendu « est désormais largement ancré dans le discours public » sur le développement, mais aussi dans la presse et le langage courant.

Imaginez quatre unités (des villages, par exemple), dont deux seront traitées et deux autres non. Pour ce faire, on peut demander aux anciens du village de décider, par exemple par appel d'offres (ou suite à un pot-de-vin), quels villages sont inclus (ou exclus), puis de sélectionner en vue du traitement les deux villages qui tiennent le plus à être traités (ou y sont le moins réticents). L'attribution des traitements et des contrôles par autosélection est clairement problématique. Beaucoup semblent penser que la randomisation règle le problème de l'autosélection. Il n'y a que six attributions possibles, dont l'une est l'autosélection. On se retrouve donc dans la situation absurde ou une même attribution est jugée satisfaisante si elle a été établie par randomisation, mais ne l'est plus si elle a été déterminée par autosélection. Sur des centaines de villages, l'équilibre ou non dépend du nombre de facteurs à équilibrer, et rien n'empêche que l'attribution réelle soit justement l'attribution par autosélection que nous voulions pourtant éviter. Rien n'est *garanti* par la randomisation. C'est peut-être l'idée selon laquelle la randomisation est équitable *ex ante* qui amène

certains à penser, par confusion, qu'elle l'est aussi *ex post*. Mais c'est l'*ex post* qui compte.

Faire en sorte que les groupes de traitement et de contrôle se ressemblent est une bonne chose, mais cela nécessite des informations et une attribution *réfléchie*, ce qui n'est pas compatible avec la randomisation. Fisher en était conscient et savait qu'il existait des moyens plus précis d'estimer un effet de traitement moyen en évitant la randomisation. Toutefois, il avait compris qu'il n'était pas facile de savoir quoi penser de la différence une fois celle-ci mesurée ; il existera toujours une *certaine* différence, même lorsque le traitement n'a aucun effet pour aucune unité. La randomisation apporte une solution à ce problème, car elle permet d'émettre des hypothèses probabilistes sur le fait que la différence soit le fruit du hasard ou non. Il y a de nombreuses années, le philosophe SUPPES (1982) l'avait formulé de cette manière. Il s'est imaginé qu'on lui présentait une urne avec 50 boules noires et blanches ; il y a soit (A) 15 boules noires et 35 boules blanches, soit (B) 35 boules noires et 15 boules blanches. Il est autorisé à tirer 12 boules et doit parier sur A ou B. Il avait écrit : « J'ai du mal à imaginer un parieur averti qui n'exigerait pas une telle randomisation physique avant de participer à l'expérience ». La randomisation *ne garantit pas* l'équilibre, mais *elle permet* de calculer les probabilités, du moins dans des cas simples comme celui-ci, où rien d'autre n'affecte les résultats. Le calcul des probabilités est utile et important, mais ce n'est pas la même chose que l'équilibre.

Généralement, les gens sont surpris lorsqu'on dit que l'inférence sur une moyenne – et donc l'inférence sur la différence entre deux moyennes – constitue un problème non résolu. BAHADUR et SAVAGE (1956) ont déjà soulevé la question il y a longtemps, en montrant que sans hypothèses limitant l'asymétrie (*skewness*), la valeur de *t* calculée ne suivra généralement pas la loi de Student. Si nous supposons à tort que c'est le cas, nous commettrons des erreurs, par exemple en pensant qu'une valeur *t* élevée indique un effet du traitement, alors qu'en fait, il n'y en a pas. L'asymétrie (terme souvent utilisé à tort de nos jours pour désigner le biais) fait référence au moment d'ordre trois, et notamment à la présence de valeurs aberrantes d'un côté de la distribution. On peut imaginer un tel problème dans toute expérimentation impliquant de l'argent ou dans le domaine de l'éducation ou de la microfinance, où une ou deux personnes sont extrêmement douées, et les autres le sont moins (BANERJEE *et al.*, 2019a).

Dans le cadre de l'expérience de la RAND sur la santé – l'une des plus célèbres RCT dans le domaine de l'économie – une participante avait eu une grossesse extrêmement coûteuse. Dans ce type de cas, le résultat de la RCT dépend du fait que la ou les valeurs aberrantes se trouvent dans le groupe de traitement ou dans le groupe de contrôle, et avec une valeur aberrante « extrême », soit peu de choses. Vous pensez peut-être disposer de centaines ou de milliers d'observations, mais en fait, vous n'en avez qu'une seule. Les valeurs aberrantes passent pour significatives, car l'utilisation de la distribution *t* est invalidée par l'asymétrie. Tronquer les valeurs aberrantes ou transformer la variable de résultat, par exemple en prenant des logarithmes, n'aideront pas toujours. Le « bébé à un million de dollars » va provoquer la faillite d'un régime d'assurance dans le

monde réel, même si les assureurs préfèrent « exclure une valeur extrême ». Nous devons mesurer les bénéfices en dollars, et non en logarithmes de dollars, sans parler des dollars « après exclusion des valeurs extrêmes ». L'effet médian du traitement serait peut-être plus fiable, mais, une fois encore, c'est la moyenne qui rend le budget déficitaire, pas la médiane, et même dans les cas où nous voudrions connaître l'effet *médian* du traitement, il n'est pas possible de l'obtenir à partir d'une RCT. Si vous vous intéressez vraiment à la médiane, vous devrez utiliser une autre méthode que les RCT, une méthode qui requiert davantage d'hypothèses.

Le problème ici, ce n'est pas que les RCT rencontrent des difficultés particulières, c'est qu'elles ne sont pas exemptes de ces difficultés. Ulrich Mueller a récemment montré que le problème est très répandu dans l'économie appliquée contemporaine, notamment lorsqu'on utilise des erreurs types robustes par *cluster* (MUELLER, 2019). Lorsque les *clusters* sont de tailles différentes – comme dans beaucoup de travaux d'économétrie appliquée avec une dimension spatiale –, les valeurs *p* qui proviennent de Stata[3], par exemple, ne sont pas fiables. Je pense que le travail de Mueller, qui fournit également une meilleure méthode, modifiera considérablement notre façon de travailler et ce que nous pensons savoir.

Dans ses travaux sur un autre problème lié à l'inférence, Alwyn Young a démontré que de nombreuses publications basées sur des RCT n'ont pas les bonnes valeurs *p* (YOUNG, 2019), de sorte que de nombreux résultats apparemment significatifs – et parfois assez surprenants – sont le fruit du hasard en l'absence d'un effet du traitement. Young propose de revenir à la randomisation à la Fisher pour calculer la significativité. Si le traitement n'a d'effet pour personne et qu'il n'y a pas de facteurs de confusion (*confounding factors*) post-randomisation, l'effet moyen estimé du traitement n'est que le résultat de l'assignation aléatoire des sujets au groupe du traitement ou au groupe de contrôle. Les facteurs de confusion post-randomisation comprennent tous les éléments qui affectent les résultats, mais qui ne relèvent pas du traitement, comme le bouche-à-oreille dans les zones de traitement, où le programme est proposé, ou le fait que l'assignation au groupe de traitement ou de contrôle ne se fasse pas à l'insu des sujets, des évaluateurs ou des analystes. En examinant toutes les répartitions aléatoires possibles dans les données réelles, nous pouvons établir la distribution des différences entre les deux moyennes sur la base de l'hypothèse d'une absence d'effet pour chaque individu, et calculer ainsi la probabilité d'obtenir un résultat aussi important ou plus extrême que la différence réelle. Cette « inférence de randomisation » teste l'hypothèse selon laquelle le traitement n'a aucun effet pour *personne*. Si cette hypothèse est généralement intéressante, elle ne nous aide pas vraiment à déterminer ce qui nous intéresse souvent en matière de politique, à savoir si l'effet *moyen* du traitement est nul. Alors qu'un effet nul pour chaque observation signifie nécessairement que la moyenne doit également être nulle, l'inverse n'est pas vrai, notamment lorsque le traitement affecte de

3. Stata est l'un des principaux logiciels de traitement des données statistiques.

manière contraire différents individus. Un exemple parlant est celui de l'aspirine : une petite dose quotidienne en sauve certains et en tue d'autres. En matière de politique publique, par exemple dans le cadre d'une expérience dans le domaine de l'enseignement, ce qui nous intéresse, c'est de savoir si la nouvelle méthode améliore les notes obtenues aux examens en moyenne, et pas seulement si elle est efficace pour une personne. Autre aspect complexe : il arrive qu'un test statistique admette l'hypothèse selon laquelle chacune des estimations dans un groupe est égale à zéro, mais rejette l'hypothèse selon laquelle leur moyenne est égale à zéro. Par ailleurs, « l'inférence de randomisation » peut elle-même être faussée par un mauvais échantillon.

Les niveaux de significativité calculés n'étant pas fiables dans des situations réalistes, il convient de lire avec prudence bon nombre des conclusions de RCT publiées. *Poor Economics* (Banerjee et Duflo, 2011) présente les conclusions de dizaines d'études, dont beaucoup sont intéressantes et importantes. Mais des résultats qui devraient être annoncés comme des estimations ont tendance à être présentés comme s'il s'agissait de faits établis. En effet, selon la rhétorique des RCT, les essais peuvent établir la vérité. Or, ce n'est pas le cas. Parfois, certains résultats surprenants issus de RCT ne sont absolument pas des résultats, et les valeurs *t* élevées ne devraient pas nous faire croire le contraire.

Les RCT sont rigoureuses et scientifiques

Ces adjectifs sont souvent utilisés de façon systématique pour qualifier les RCT, et pourtant cette affirmation est rarement, voire jamais, justifiée. Or, cette rhétorique semble avoir du succès, du moins auprès des bailleurs de fonds. Elle va souvent de pair avec un plaidoyer en faveur de l'importance des RCT en médecine, mais rarement avec une analyse réaliste des succès et des échecs des RCT en médecine. Aux États-Unis, les médicaments doivent faire l'objet de RCT concluantes afin d'être autorisés. Pourtant, les opioïdes sur ordonnance, tels que l'OxyContin, ont tué des centaines de milliers d'Américains au cours des vingt dernières années. Les RCT ne fonctionnent pas de la même façon en sciences sociales et en médecine, sujet qui mérite une réflexion plus approfondie. Un jour, j'ai discuté avec un responsable du financement d'une grande fondation d'une série d'essais randomisés dans le champ du développement. Il a volontiers admis que l'applicabilité des résultats était limitée et que certains d'entre eux étaient sans doute incorrects, mais cela ne l'a pas gêné pour autant. Après tout, m'avait-il dit, les RCT sont plus rigoureuses que toute autre méthode. Pour lui, cet argument suffisait. Il avait, je pense, la conviction que rigueur était synonyme de résultats généralisables. Ou peut-être était-il persuadé, comme beaucoup, que toutes les autres méthodes sont pires. Pour lui, se tromper ne semblait pas incompatible avec le principe de rigueur.

Validité externe

« Trouver ce qui fonctionne » est un autre grand slogan rhétorique qui, à en juger par le nombre de fois où il est répété, fait son effet auprès du public. Pourtant, rien ne fonctionne sauf dans un certain contexte, c'est pourquoi trouver ce qui fonctionne, où et dans quelles circonstances, est une véritable mission scientifique. Ce qui fonctionne dépend aussi des bénéficiaires et de l'objectif visé ; cela implique des valeurs aussi bien que des faits. Il n'existe aucune expérimentation ou série d'expérimentations qui puisse répondre sans réserve à ces questions. Dire que les RCT permettent d'identifier ce qui agit pour éliminer la pauvreté dans le monde est tout à fait louable, mais c'est une prétention infondée.

Un résultat qui s'avère valable à un endroit précis, à un moment donné et dans un ensemble de circonstances, ne le sera généralement pas ailleurs, à un autre moment ou dans des circonstances différentes. Ce qui a du succès pour vous peut « marcher » pour moi aussi, mais cela peut ne pas me plaire. Encore une fois, cela vaut pour tous les résultats empiriques, quelle que soit la méthode utilisée. Une estimation du revenu moyen aux États-Unis ne sera plus valable dans dix ans, personne ne vous dira le contraire. Pourtant, il est courant de considérer que l'estimation d'un effet de traitement moyen, qui est également une estimation d'une moyenne basée sur un échantillon, est susceptible de se vérifier ailleurs, sauf preuve du contraire.

Cette pratique diffère sans doute peu d'une habitude bien ancrée en économie qui consiste à aborder les élasticités comme des constantes : on parle de « l'élasticité » de la main-d'œuvre masculine située dans la tranche d'âge de forte activité ou de « l'élasticité prix » du pain. Je pense que ces élasticités s'appuient sur de fortes intuitions quant à la nature des biens concernés : que la plupart des hommes n'avaient guère d'autre choix que de travailler, alors qu'autrefois leurs épouses l'avaient davantage, que les aliments de base ne sont pas faciles à remplacer, et que la demande de « biens de luxe du quotidien » est influencée par leurs prix, intuitions qui ont été confirmées par de nombreuses études menées en divers endroits. Mais le développement n'en est pas là actuellement. Pour reprendre l'exemple de Pritchett (chap. 2, ce volume), il n'y a, à mon sens, aucune raison de supposer que, s'il vaut mieux des poulets que de l'argent en Sierra Leone, ce sera aussi le cas au Laos ou encore à Trenton, dans le New Jersey. Aucune raison de penser non plus que, s'il valait mieux des poulets dans 60 essais menés dans 60 lieux différents, ce serait aussi le cas dans le 61e. Et n'oubliez pas la poule de Bertrand Russell, qui, après avoir vécu des centaines de fois le même scénario, savait que les pas de l'agriculteur signalaient le moment où elle allait être nourrie, jusqu'à ce que, la veille de Noël, il lui torde le cou. Comme l'a fait remarquer Bertrand Russell, il eût été bien utile audit poulet d'avoir une vision plus subtile de l'uniformité de la nature (Russell, 2012).

Une meilleure compréhension est essentielle. La fondation Gates, le plus grand bailleur de fonds dans de nombreux domaines, considère la mise à l'échelle comme l'une de ses missions centrales. Elle a donc pris un ou deux résultats

positifs de son initiative pour l'agriculture africaine pour prouver que « ça fonctionne », et l'a étendue à d'autres exploitations agricoles et à d'autres pays, sans chercher à savoir si cela pouvait ou non agir efficacement ailleurs (SCHURMAN, 2018). Il faut se rendre à l'évidence : ce qui a du succès dans une ferme ne fonctionne pas forcément dans une autre, constat que les agriculteurs africains ont sans doute fait, contrairement aux expérimentateurs.

C'est une erreur de penser que validité interne et validité externe vont systématiquement de pair et qu'elles caractérisent les études de « grande qualité ». Une RCT peut être parfaitement menée sur un large échantillon et l'*Average Treatment Effect* (ATE) calculé de façon précise. La validité externe n'est *pas* une caractéristique de *l'étude*, mais une caractéristique des *circonstances* dans lesquelles elle sera mobilisée. Une étude dont le résultat ne s'applique pas ailleurs n'est en rien invalide. La validité externe porte sur la manière dont une étude est *mobilisée* ; une même étude peut être valide dans certains contextes et pas dans d'autres.

Il est toujours tentant de prendre une étude impressionnante et de la porter au-delà de son contexte d'origine. Cela vaut aussi bien pour les études non expérimentales que pour les études expérimentales. Raj Chetty et ses coauteurs (CHETTY *et al.*, 2019) ont été des pionniers de l'utilisation des données administratives fusionnées pour décrire de manière incroyablement détaillée des faits sur la dynamique des inégalités aux États-Unis, et ont ainsi permis des avancées considérables en matière de connaissances. Selon leur conclusion importante, entre 1989 et 2015, les enfants afro-américains avaient une plus faible probabilité que les enfants blancs d'avoir des revenus supérieurs à ceux de leurs parents. Pourtant, dans de nombreux articles de presse, l'imparfait « avaient » est remplacé par le présent « ont », et ce malgré les évolutions des tendances relatives au mariage et à l'incarcération dans les deux groupes. Ces études sont certes éminentes, parmi les meilleures de l'économie d'aujourd'hui, mais elles ne sauraient prétendre à une meilleure validité externe que les RCT les plus remarquables. Une fois de plus, la question de la validité externe a une portée générale et les RCT n'y échappent pas. S'il est vrai que sans validité interne, le résultat d'un essai a peu de chances d'être concluant ailleurs, on ne peut affirmer que la validité interne entraîne nécessairement la validité externe. Je n'ai jamais entendu dire le contraire, mais je suis souvent frappé par le contraste entre le soin apporté à la réalisation d'une RCT et l'insouciance dont on fait preuve en préconisant l'utilisation de ses résultats. La formule selon laquelle « la validité interne prime » semble servir à justifier de telles pratiques.

Le fait d'imaginer que les résultats d'une RCT seront utilisés dans un contexte différent de celui dans lequel elle a été réalisée peut contribuer à orienter la conception initiale de l'essai afin de le rendre plus utile. Si nous partons du principe que les effets du traitement sont différents selon les sous-populations, alors une stratification selon ces sous-populations permettra non seulement d'améliorer la précision de l'essai, mais aussi de procéder à une nouvelle pondération pour l'adapter à une nouvelle situation. La mise à l'échelle affectera souvent des variables potentielles qui s'avèrent être des constantes entre les branches de l'essai. Par exemple, si dans le cadre d'une politique éducative,

davantage d'étudiants sont formés, les salaires risquent de baisser, de sorte que prévoir dans l'essai une branche sur les salaires bas pourrait apporter des informations utiles. Les RCT peuvent contribuer à fournir les outils permettant de modéliser les conséquences politiques plutôt que simplement de contourner ou ignorer le fossé qui existe entre un essai et sa mise en œuvre. Mais une RCT ne suffira probablement pas à elle seule.

Le fait que les résultats d'une même étude soient reproduits dans des contextes et des pays différents – comme c'est le cas de l'étude sur les programmes de formation (BANERJEE *et al.*, 2015a) parue dans *Science* – est effectivement surprenant. Cependant, on n'est pas certain que les gains puissent être reproduits par des fonctionnaires confrontés à des incitations financières et politiques réalistes, très différentes de celles auxquelles sont confrontés les assistants diplômés étrangers hautement qualifiés, désireux de voir leur projet réussir. Cette étude transnationale met par ailleurs en avant le manque de clarté quant à la réplication : que signifie-t-elle ? Quelle mesure voulons-nous répliquer ? Que pouvons-nous apprendre de la réplication ? Nous pourrions chercher à obtenir le taux de retour sur investissement, ou bien la proportion de personnes ayant dépassé un seuil de pauvreté local ou mondial par unité de monnaie internationale. Les auteurs utilisent plutôt la « taille de l'effet », qui est l'ATE standardisé par l'écart-type du traitement. Selon GOLDBERGER et MANSKI (1995 : 769), « la standardisation se borne à donner aux quantités en unités non comparables l'apparence superficielle d'être en unités comparables. Elle est donc pire qu'inutile – elle donne lieu à des inférences trompeuses ».

Pré-enregistrement des essais

L'American Economic Association (AEA) exige que les essais dont les résultats doivent être publiés dans ses revues soient soumis à un enregistrement préalable, ce que j'ai contesté, mais en vain. Je ne pense pas que l'AEA ait intérêt à légiférer sur les méthodes, elle devrait plutôt évaluer les études en fonction de leurs mérites. Mon expérience en tant qu'économiste et membre des comités de l'AEA m'a appris que les désaccords entre économistes qui sont, en vérité, politiques ou personnels, sont souvent présentés comme des divergences méthodologiques. L'AEA a réussi, au moins depuis les années 1930, à éviter tout schisme ; elle est restée une communauté très large, intégrant des économistes de tous bords, présidée par des personnalités allant de Milton Friedman à Kenneth Galbraith, même si je doute qu'ils aient beaucoup d'estime pour les méthodes des uns et des autres (Friedman ayant essayé en vain de faire obstacle à la présidence de Galbraith).

Les problèmes de *p-hacking*, de *data mining* et de recherche de spécifications sont bel et bien réels. Les bailleurs de fonds qui ont déboursé d'importantes sommes pour une RCT font généralement pression pour trouver au moins un sous-groupe pour qui le traitement s'est avéré efficace. Mais, là encore, ces

problèmes ne sont pas spécifiques aux RCT. Certains ont en effet plaidé pour un pré-enregistrement de toutes les études, de sorte qu'avant de commencer à travailler sur une étude d'observation basée sur le recensement, par exemple, je devrais informer l'AEA – ou peut-être le *Census Bureau* – de mon plan d'analyse des données. Mais il est loin d'être évident de savoir jusqu'où aller : dois-je rapporter une conversation avec un collègue ou une découverte que j'ai lue dans le journal et qui oriente mon programme ou limite mon choix de variables ?

Les découvertes dont je suis le plus fier revêtent toutes une part importante de hasard, même si j'étais suffisamment informé pour savoir ce que je regardais, même lorsque je cherchais autre chose. Aucun de ces résultats ne serait apparu dans un plan de pré-analyse et n'aurait donc pu être publié dans la *Revue des études correctement réalisées*. Bill Easterly a fait remarquer que Christophe Colomb n'aurait pas pu découvrir l'Amérique s'il avait été obligé de s'en tenir à un plan de pré-analyse déposé dans un coffre-fort à Séville ou à Gênes (Easterly, 2012). J'ai du mal à croire que ce que nous avons trouvé, Anne Case et moi, sur les taux de mortalité des personnes en milieu de vie (Case et Deaton, 2015), des résultats auxquels nous ne nous attendions absolument pas, soit attribuable à l'exploration des données (*data snooping*). Pourtant, j'imagine très bien un rédacteur en chef borné sur le plan statistique rejeter l'article parce que nous n'avons pas présenté le certificat de pré-enregistrement qui autorisait notre travail sur la mortalité des adultes en milieu de vie. Le risque d'étouffer des résultats importants, mais inattendus est sans doute bien pire que le risque de promouvoir des résultats fallacieux.

Expérimentation : donnez un coup de pied et vous verrez

Je suis tout à fait favorable à l'expérimentation (Morduch, chap. 3, ce volume). Mais il n'y a aucun lien logique entre expérimentation et randomisation. En effet, lorsque l'on s'apprête à donner un coup de pied, mieux vaut savoir précisément ce que l'on vise. Il est déconseillé de donner un coup de pied au hasard, au risque de blesser. La randomisation consiste à apprécier la signification de l'événement, et non à préparer un coup de pied. Ce qui est important à retenir ici, c'est que, bien souvent, la randomisation n'est pas utile à l'expérimentation, et peut même transformer une bonne expérience en une expérience inutile, en brouillant des informations qui devraient servir à améliorer notre étude.

Les grandes expériences de laboratoire en économie n'ont pas eu recours à la randomisation (Svorenčik, 2015). La révolution industrielle est souvent décrite comme étant le fruit de bricolages sans fin, et non de la randomisation, qui aurait empêché les essais et erreurs volontaires. Un autre exemple que j'ai déjà utilisé

dans le passé (DEATON, 2012) est le jeu vidéo d'arcade *Angry Birds* qui consiste à propulser des oiseaux à l'aide de lance-pierres. Il est possible de les rediriger, d'accélérer le tir, de les faire exploser en plein vol, le but étant de tuer les cochons voleurs d'œufs qui se cachent dans des endroits inaccessibles. Étant donné les innombrables combinaisons possibles, il faudrait énormément de temps pour réaliser une série systématique de RCT, encore qu'un enfant habile pourrait trouver la solution en quelques minutes. Il existe de nombreux types d'expérimentations où la randomisation n'est pas nécessaire ou risquerait de brouiller les résultats. Après tout, la randomisation est aléatoire. Chercher des solutions au hasard est donc inefficace parce que cela revient à envisager beaucoup de possibilités non pertinentes, comme ce fut le cas dans les champs de FISHER (1960).

RCT et autres méthodes

Lorsqu'on parle de RCT, bien souvent on les compare avec d'autres méthodes, généralement celles des variables instrumentales (VI), de régression par discontinuité (RD) ou des différences de différences. Cette comparaison est toutefois beaucoup trop étriquée. Les méthodes économétriques sont mon quotidien, je les utilise et les enseigne depuis plus de quarante ans, j'ai donc pu observer la progression qui a conduit aux RCT. Avant, nous faisions des régressions de y en x, sans trop discuter de ce qui générait la variation en x. Nous avons découvert que les méthodes des différences de différences, des variables instrumentales et de régression par discontinuité permettaient de supprimer la variance indésirable de x et de créer deux groupes jugés identiques, sauf le traitement. On pourrait voir les RCT comme des versions plus épurées des VI, RD, ou des différences de différences, revenant effectivement à la régression, mais avec l'hypothèse garantie que x a été attribué de manière aléatoire. À la lumière de cette évolution, on peut comprendre pourquoi les RCT semblent être la solution ultime, et finissent par le devenir lorsque l'on pense de la sorte.

Mais, comme l'a fait remarquer John Stuart Mill il y a déjà bien longtemps (MILL, 1843), la « méthode des différences », qui consiste à comparer deux groupes, l'un traité, l'autre non, n'est qu'une façon parmi d'autres de faire une inférence causale. Trouver la cause d'un accident d'avion n'implique pas de différence (ou du moins on l'espère), et la méthode hypothético-déductive, utilisée par les physiciens, n'implique pas de différence, mais simplement la formulation et la vérification de prédictions. C'est pourquoi, lorsqu'ils organisent les données d'une manière qui contredit bon nombre de nos connaissances préalables sur le fonctionnement du monde, les graphiques et les tableaux croisés peuvent être si puissants. Plus spécifiquement, la commission Cowles a mis au point une méthode de construction de modèles causaux en accordant une attention particulière aux mécanismes et en utilisant un langage qui met l'accent sur la structure causale et les procédures permettant de délimiter les éléments de la structure qui peuvent ou non être estimés

à partir des données. Ces modèles pouvaient être examinés afin de tester leurs prédictions et l'adéquation de la structure causale. À l'époque, les économistes utilisaient davantage ces méthodes qu'aujourd'hui, et, pendant de nombreuses années, elles ont été au cœur des textes d'économétrie. Pourtant, je pense que la plupart des étudiants en économie auraient aujourd'hui du mal à définir les formes structurelles et réduites. Les articles comportaient une partie théorique, qui développait des prédictions vérifiables, de préférence des prédictions surprenantes et propres à la théorie, et qui étaient ensuite vérifiées dans la partie empirique. Certaines de ces méthodes peuvent être interprétées comme permettant d'examiner les différences entre les groupes, mais pas toutes.

Petits *versus* grands effets

Lant Pritchett a avancé un argument particulièrement éloquent, cocasse et passionné : ce qui importe pour réduire la pauvreté, c'est la croissance, et non une évaluation « rigoureuse » (ou non) de chaque projet, qu'il s'agisse d'argent ou de poulets (Pritchett, chap. 2, ce volume). Dans *Poor Economics*, BANERJEE et DUFLO (2011) affirment le contraire : ce n'est qu'au cas par cas que nous savons ce que nous faisons, c'est pourquoi nous devons construire des connaissances par le biais de la randomisation, essai après essai.

Le débat est (au moins) aussi vieux que la Banque mondiale. Je vais vous en faire un bref historique. La Banque mondiale a commencé « petit », avec des projets de construction de ports, de routes, de centrales électriques, etc. Très vite, elle s'est rendu compte qu'évaluer des projets selon des critères commerciaux ne permettait souvent pas d'améliorer la vie des gens, notamment dans les économies où les prix sont faussés par les tarifs, les offices de commercialisation, le rationnement ou le contrôle des changes. Deux groupes d'éminents économistes n'ont pas tardé à réagir en instaurant des prix fictifs pour remplacer ceux du marché. DASGUPTA *et al.* (1972) ont mis au point une série de méthodes pour les Nations unies ; LITTLE et MIRRLEES (1974) en ont élaboré une autre pour l'OCDE. Cette dernière a fait l'objet d'un manuel élaboré par SQUIRE et VAN DER TAK (1975) à l'intention de la Banque mondiale. Les calculs étaient cependant parfois compliqués, dépassant les capacités ou la volonté des agents chargés des prêts, dont les motivations premières étaient de faire circuler l'argent rapidement. Ces règles ont dû sembler incompréhensibles aux décideurs politiques des pays auxquels on a demandé de les appliquer. Pour illustrer l'état primitif de l'évaluation des projets dans une grande partie du monde, SQUIRE a noté plus tard (1989 : 1126-1127) que même l'outil d'évaluation de projet le plus basique, à savoir l'actualisation des bénéfices futurs, était rarement utilisé dans les pays emprunteurs livrés à eux-mêmes. Ce n'était pas le cas en Inde, où les économistes de la Commission de planification calculaient méticuleusement des prix fictifs, certains ravalant néanmoins leur scepticisme. Si l'économie

était complètement faussée, évaluer les projets selon les prix du marché ne servait certainement à rien, et les évaluer selon des prix fictifs n'était pas une alternative possible.

La solution consistait à passer du petit au grand, à corriger les distorsions en priorité et à remettre la macro-économie sur des rails avant de procéder à l'évaluation des projets, avec comme résultat l'ajustement structurel.

Pour soutenir cela, des analyses empiriques, comme celle de Pritchett, ont montré que la croissance économique pouvait contribuer à une réduction importante de la pauvreté. Les grands épisodes de réduction de la pauvreté dans le monde – notamment en Chine et en Inde – ont été alimentés par la croissance économique et par la mondialisation. La croissance globale a été alimentée par la croissance à petite échelle, avec davantage d'emplois, d'opportunités, de routes, d'écoles et de cliniques de meilleure qualité, mais tous ces éléments étaient perçus comme des phénomènes plus ou moins spontanés dans une économie dotée de bonnes institutions et en pleine croissance. Rien de tout cela ne permettait d'expliquer comment stimuler la croissance économique. Les régressions où les unités d'analyse sont les pays ont donc été considérées comme bénéfiques. Elles ont fait l'objet de nombreuses critiques et de railleries, mais elles ont permis d'acquérir de précieuses connaissances, notamment sur l'importance de l'investissement intérieur – essentiel en Chine, en Inde et en Corée – pour la fourniture de biens publics, et sur le fait que l'aide étrangère, aussi bénéfique soit-elle, ne peut guère stimuler la croissance à elle seule. Elles ont par ailleurs systématisé et structuré les résultats, ce qui était préférable aux anecdotes auxquelles chaque pays se livrait individuellement et qui avaient dominé une grande partie de la précédente discussion. Mais nous en avons appris davantage sur ce qui ralentit la croissance que sur ce qui l'accélère. Tous ces éléments sont précieux, mais ne constituent guère la clé pour éliminer la pauvreté en accélérant la croissance. Personne, à ma connaissance, n'a laissé entendre que les RCT étaient la clé de la croissance économique. Difficile d'ailleurs de trouver un exemple où les RCT ont eu un quelconque impact sur la réduction de la pauvreté en Chine (Yang, 2019).

La Banque avait à moitié raison. Une meilleure gestion macro-économique dans beaucoup de pays du monde et une meilleure compréhension de la politique monétaire et du rôle de la banque centrale, ainsi que des effets de la sous-évaluation des taux de change et de la taxation des prix des matières premières, ont contribué à une croissance plus forte et une réduction de la pauvreté, surtout avec le temps (Easterly, 2019). Les économistes d'aujourd'hui, partisans de la révolution de la crédibilité et des tests de causalité, ont tendance à rejeter ces résultats arguant qu'ils ne sont, selon eux, ni rigoureux ni crédibles. Pourtant, ils n'ont aucun mal à affirmer que les RCT permettent de réduire efficacement la pauvreté dans le monde.

Ceux qui pensent que l'aide extérieure peut favoriser le développement économique doivent chercher la quadrature du cercle. Personne ne doute de l'importance de la perspective macro, mais il faut reconnaître que les outils permettant

d'influencer la croissance économique sont limités. Les essais au niveau micro s'avèrent souvent fructueux, mais leur capacité à réduire les taux de pauvreté relève essentiellement d'une question de foi. Les RCT ne peuvent se passer d'une théorie de mise en œuvre, ou de mise à l'échelle, pour guider l'exploitation des résultats dans la pratique. Il convient notamment de prêter attention aux conséquences involontaires – les effets de la mise en œuvre sur les actions du gouvernement et les communautés – qui ne figurent généralement pas dans les critères d'évaluation de l'essai randomisé. Il faut réfléchir aux effets d'équilibre général, car la mise à l'échelle modifiera des prix et des comportements qui étaient constants dans les expérimentations. Lors d'une RCT, on part souvent du principe que les effets indirects n'existent pas (hypothèse dite *Stable Unit Treatment Value Assumption* – SUTVA). Or, cette hypothèse est régulièrement enfreinte, par exemple dans les projets d'assainissement (Spears, Ban et Cumming, chap. 6, ce volume) ou de déparasitage. Sur le plan individuel, les effets du traitement et les effets indirects sur les autres sont faibles et peuvent rarement être mesurés (ou ne le sont pas). Pourtant, au niveau global, la somme des petits effets indirects et individuels peut annuler ou inverser l'effet.

Modèles

Il existe une tentation très forte de chercher à faire des recommandations politiques sans avoir à construire des modèles. Je comprends qu'il puisse être tentant de laisser parler les données, ou de générer des données qui parlent d'elles-mêmes, mais je crois que ces tentatives sont vouées à l'échec. Pour interpréter une RCT, il faut toujours faire des hypothèses. Nous devons partir du principe que seul le traitement importe, ce qui est impossible à garantir sans une surveillance attentive des facteurs de confusion post-randomisation, tout comme il est impossible d'être sûr que les restrictions d'exclusion sont valables pour faire une estimation à l'aide de variables instrumentales. Les sujets n'acceptent pas toujours leur attribution, qui peut être gérée grâce à l'estimation de l'intention de traiter (*intent to treat*), même si, bien souvent, ce n'est pas l'effet moyen que l'on s'emploie à découvrir. Nous pouvons aussi construire des modèles expliquant pourquoi les gens acceptent ou n'acceptent pas leur attribution, ce qui est en soi une information potentiellement utile (Heckman et Smith, 1998). Que se passe-t-il si une RCT donne un effet positif lorsque le résultat est mesuré en niveaux, mais qu'elle aboutit à un effet nul lorsque ledit résultat est mesuré en logarithmes ? Ces cas sont faciles à construire[4].

4. Imaginons un traitement binaire simple, qui modifie le revenu logarithmique d'un montant a variant selon les unités, mais dont la moyenne est égale à zéro. L'effet sur le revenu individuel est $a.y$ (y étant le revenu). La moyenne de $a.y$ dépend de la corrélation entre le revenu et l'effet du traitement individuel, qui peut être positif, négatif ou nul.

Comme les praticiens le savent, l'utilisation d'informations préalables permettra d'améliorer la précision (Vivalt, chap. 11, ce volume). En pratique, les effets moyens des traitements sont souvent estimés en effectuant une régression qui inclut des variables de contrôle. Celles-ci doivent être choisies, mais on ne sait pas exactement selon quelles règles les variables sont incluses ou exclues, ni combien en utiliser. La stratification peut également accroître la précision, mais uniquement si elle utilise des informations préalables valables sur les différences des effets moyens du traitement entre les strates.

Pour *exploiter* les résultats des essais, la modélisation est essentielle. Nous avons besoin d'une théorie sur laquelle nous appuyer pour savoir si les résultats sont pertinents ailleurs et, si oui, comment les adapter.

Causalité

Une RCT bien conçue nous renseigne sur la causalité. Mais, là encore, de nombreuses hypothèses doivent être formulées pour passer des données à la conclusion. Dans tout essai randomisé en univers fini, et ils le sont tous, on ne peut jamais exclure la possibilité que le résultat soit dû au hasard. La mesure des résultats peut avoir son importance, comme dans l'exemple sur les niveaux *versus* les logarithmes. Pour citer les philosophes et épidémiologistes Broadbent *et al.* (2017 : 1844), « Les conclusions causales ne découlent pas de manière déductive des données sans un solide ensemble d'hypothèses auxiliaires, et ces hypothèses ne sont elles-mêmes pas des conséquences déductives des données. » Dans le même article, ils écrivent : « Selon nous, il est préférable de ne pas qualifier l'estimation d'une étude individuelle comme étant "causale", même s'il s'agit d'un essai randomisé. C'est l'ensemble des preuves qui permet de déterminer la causalité. La causalité est une conclusion scientifique, une affirmation *théorique*, et en tant que telle, elle transcende toute étude individuelle » (italique ajouté). La causalité relève de l'esprit, pas des données, une idée que Heckman et Pinto attribuent à Frisch et Haavelmo (Heckman et Pinto, 2015). La triangulation des résultats, ou la découverte des processus de causalité résultant de nombreuses études menées dans le temps est bien illustrée dans cet ouvrage par le chapitre sur l'assainissement (Spears, Ban et Cumming, chap. 6, ce volume).

Il convient de noter que ce ne sont pas seulement les résultats d'une RCT qui peuvent ne pas être transférables, mais la causalité elle-même. Cartwright et Hardie (2012) illustrent cela avec une machine de Rube Goldberg, où l'ouverture d'une fenêtre mène, à l'issue d'une longue chaîne de liens de causalité absurdes, mais efficaces, à un crayon taillé par un pivert. Pourtant, ce n'est généralement pas en ouvrant les fenêtres que l'on taille des crayons, de même qu'une chaîne causale dans un contexte donné peut être très différente dans un autre contexte. J'ai l'impression que, lorsque les économistes utilisent le mot « causal » dans

les titres de leurs articles, ils invoquent plutôt un seul cas dans un contexte spécifique. Méfiez-vous de Rube Goldberg.

Les étudiants en économie formés dans la tradition de Cowles, ainsi que les lecteurs de Judea Pearl (PEARL et MACKENZIE, 2018) savent qu'il existe d'autres façons de construire des modèles de causalité. Pearl soutient que nous devons *commencer* par un modèle causal et l'utiliser ensuite pour confronter les données et tester sa structure et, comme la commission Cowles l'a fait avant lui, il propose une série d'outils et de méthodes pour y parvenir. BRADFORD-HILL (1965) a étudié avec finesse les nombreux moyens de détecter la causalité, mais il semble que l'économie s'y réfère peu. Bradford-Hill était le pionnier des essais cliniques randomisés il y a soixante-dix ans et il semble parfois que nous perdons des connaissances plutôt que d'en acquérir.

Éthique

Les économistes doivent bien réfléchir à l'éthique des expérimentations. Je n'ai pas grand-chose à ajouter aux discussions sur l'équipoise et le consentement éclairé, qui sont traitées ailleurs (Abramowicz et Szafarz, chap. 10, ce volume). Pourtant, certaines RCT dans le domaine du développement semblent remettre en cause les règles les plus élémentaires. Que dire du consentement éclairé lorsque les personnes ne savent même pas qu'elles participent à une expérimentation ? La bienfaisance est l'une des exigences fondamentales lorsque l'on mène des expérimentations sur des sujets humains. Mais la bienfaisance pour qui ? Les expérimentateurs étrangers ou même les fonctionnaires locaux ne savent pas toujours ce que veulent les gens. Penser savoir ce qui est bon pour les autres ne suffit pas à démontrer la bienfaisance.

L'éthique exige également que nous soyons conscients de ce que les RCT sont capables de faire ou non. Les manquements à l'éthique se justifient plus facilement pour ceux qui défendent le point de vue hiérarchique, selon lequel la seule preuve qui compte est celle des RCT, excluant ainsi les options qui pourraient présenter moins de risques pour les sujets ou conduire à de meilleures conclusions. Il n'est pas éthique de dire aux responsables politiques des pays en développement que les RCT sont le seul moyen de recueillir des preuves à des fins de politiques publiques, car cela peut les amener à négliger des informations importantes. La question de l'exactitude des valeurs *p*, déjà abordée précédemment, se pose également ici. Un essai peu performant qui ne parvient pas à établir ses objectifs est également contraire à l'éthique dès lors qu'il impose des charges aux sujets.

Ma principale préoccupation va bien au-delà. Même aux États-Unis, presque toutes les RCT sur le système de protection sociale sont effectuées *par* des blancs ayant un niveau de revenus et d'éducation plutôt élevé, *sur* des personnes de couleur dont les revenus et le niveau d'éducation sont plus faibles. D'après

ma connaissance de la littérature, la plupart des expériences américaines n'ont pas été faites dans l'intérêt des pauvres qui en étaient les sujets, mais dans l'intérêt des riches (ou du moins des contribuables) qui avaient accepté, parfois à contrecœur, l'obligation de lutter contre les pires effets de la pauvreté, et qui voulaient en minimiser le coût (GUERON et ROLSTON, 2013). C'est regrettable, mais au moins, les ménages pauvres ont accès aux urnes et font partie de la société dans laquelle vivent les contribuables, avec son système de sécurité sociale. Il y a donc un retour vers leurs bienfaiteurs. Ce n'est pas le cas dans le domaine du développement économique, où les personnes aidées n'ont aucune influence sur les bailleurs de fonds. Certaines RCT réalisées par des économistes occidentaux sur des personnes extrêmement pauvres en Inde, et qui ont été approuvées par des comités d'examen institutionnels américains, semblent contraires à l'éthique, parfois même à la limite de l'illégalité, et n'auraient probablement pas pu être réalisées sur des sujets américains (SARIN, 2019). Il est particulièrement inquiétant que certaines études portent sur des questions économiques qui semblent ne présenter aucun bénéfice pour les sujets. Se servir des pauvres pour étoffer son CV ne devrait pas être admis. Aux États-Unis, alors que les comités d'examen des institutions offrent une protection spéciale aux prisonniers, dont l'autonomie est compromise, il ne semble pas y avoir de protection similaire pour certaines des personnes les plus pauvres du monde. Cela fait d'ailleurs curieusement écho aux débats sur les géants pharmaceutiques qui testent leurs médicaments en Afrique.

À mes yeux, les RCT relèvent de ce qu'EASTERLY (2013) appelle « l'illusion technocratique », soit le péché originel du développement économique et un aspect de ce que SCOTT (1998) a baptisé le « haut modernisme », selon lequel le savoir technique, même en l'absence d'une pleine participation démocratique, peut résoudre les problèmes sociaux. Selon cette doctrine, qui semble particulièrement répandue dans la Silicon Valley, parmi les fondations, et au sein du mouvement de l'altruisme efficace, la pauvreté mondiale cédera aux bonnes solutions techniques, notamment à l'adoption des RCT comme fondement des politiques basées sur les faits. Ignorer la politique est considéré à tort comme une vertu, et non comme un vice. Si les fondations et les altruistes prétendent savoir ce qui est bon pour les pauvres et sont animés de bonnes intentions, ils ne sont guère capables de prouver que les pauvres sont d'accord avec leurs évaluations ou apprécient leurs remèdes, si bien que leurs intérêts peuvent facilement entrer en conflit avec ceux qu'ils tentent d'aider. Les technocrates croient pouvoir développer les pays des autres depuis l'extérieur, parce qu'ils savent comment trouver ce qui marche. En cela, au moins, il n'y a pas de grande différence entre concevoir un gadget et concevoir une politique sociale. L'un et l'autre sont du ressort des ingénieurs.

Réduire la pauvreté par le biais de l'ingénierie est, au mieux, sans espoir et, au pire, désastreux. Les agences de développement sont très friandes du mot « partenariat », mais il n'y a pas de véritable partenariat lorsque tout l'argent se trouve d'un côté. Il ne peut pas non plus y avoir de véritable consentement éclairé dans une RCT lorsque l'argent qui est en jeu provient de l'aide au développement.

Trouver ce qui fonctionne et trouver ce qui est souhaitable sont deux choses différentes. Les intentions des bailleurs de fonds, aussi bonnes soient-elles, ne sont pas forcément le gage de ce qui est souhaité. Drèze (2018a) livre une excellente réflexion sur la question du passage de la preuve à la politique. Parmi ses exemples, on peut citer celui de la distribution d'œufs à des écoliers indiens, pays gravement touché par la malnutrition infantile. La mise en place d'une RCT pourrait permettre d'établir que les enfants qui reçoivent des œufs viennent plus souvent à l'école, étudient davantage et sont mieux nourris. Pour de nombreux bailleurs de fonds et défenseurs des RCT, cela suffirait à faire pression pour obtenir le lancement d'une politique de « distribution d'œufs aux écoliers ». Mais cette politique devra faire face à bien des réactions : l'opposition du puissant lobby végétarien, la pression de l'industrie avicole et celle d'un autre groupe qui affirmera que ses œufs en poudre – ou même son substitut d'œuf breveté – seront encore plus efficaces. Ces questions ne sont pas du ressort des expérimentateurs, mais des politiciens ainsi que des nombreux autres experts en administration politique. La plomberie du social devrait rester entre les mains des « plombiers du social », et non d'économistes expérimentaux qui n'ont aucune connaissance particulière, ni aucune légitimité (Duflo, 2017).

Travailler au service des citoyens d'autres pays comporte son lot de difficultés. Dans les pays dirigés par des gouvernements peu soucieux du bien-être de leurs citoyens – des régimes basés sur l'exploitation qui voient leurs citoyens comme une ressource à piller –, le pouvoir en place, s'il a le contrôle total, sera nécessairement le bénéficiaire de l'aide étrangère. Cela est d'autant plus flagrant dans les zones de guerre où il est impossible d'intervenir sans glisser un billet aux belligérants au risque de prolonger ou aggraver les souffrances (de Waal, 1997). Ce dilemme vaut aussi en temps de paix. Dans les régimes autoritaires qui exercent un contrôle total, l'aide étrangère n'est acceptée par le gouvernement que si elle sert son propre intérêt. Les agences de développement se voient alors « autorisées » à aider les pauvres ou à fournir des services de santé, tout en offrant une couverture politique au despote « éclairé » qui est ainsi libre de persécuter ou d'éliminer ses opposants (Deaton, 2015). Des questions similaires se posent également dans les démocraties, mais de manière moins marquée ; le passage de l'administration de la preuve à la politique n'est jamais totalement neutre d'un point de vue éthique, mais il est moins délicat lorsque les pauvres ont leur mot à dire et un certain pouvoir.

Quel est le rapport avec les RCT ? L'inutilité, tout d'abord. Il est complètement insensé de dépenser des ressources dans des essais randomisés portant sur des écoles ou des médicaments alors qu'un président, en pleine campagne électorale, est en train d'emprisonner ses opposants ou d'inciter à la violence contre ses ennemis politiques et ethniques (Wrong, 2009). Alors que la plupart des pauvres de la planète vivent dans des États soi-disant démocratiques dirigés par des autocrates populistes, les experts en essais randomisés seront confrontés à de plus en plus de dilemmes éthiques. Pourquoi les agences financent-elles des programmes d'aide, ou des RCT pour les soutenir, dans des pays dont les dirigeants ne souscrivent pas aux convictions démocratiques libérales des bailleurs

de fonds et des expérimentateurs ? Je ne dis pas qu'il n'y a pas de réponse à cette question, mais seulement que les bailleurs de fonds doivent les connaître.

La fondation Bill et Melinda Gates s'est déjà attiré les foudres de la presse[5] en remettant l'un de ses prix Global Goal à Narendra Modi pour la construction de toilettes en Inde, alors que ce même Narendra Modi prive les Cachemiris de leurs droits, menace de retirer la citoyenneté à des millions d'Assamais et utilise volontiers le critère religieux pour accorder la citoyenneté aux immigrants. La fondation soutient que la récompense ne reconnaît que les réalisations de Modi en matière d'assainissement. Cela illustre parfaitement les limites et les dangers de l'aide technocratique. Elle donne du pouvoir au despotisme et à l'intolérance. Modi a reçu d'autres prix prestigieux d'agences de développement, dont les Nations unies. Et en Afrique, il y a eu des cas bien pires.

Les agences d'aide ferment les yeux sur la répression politique tant que les oppresseurs participent à la réalisation de l'un des objectifs de développement durable, de préférence avec des essais contrôlés randomisés à l'appui. Les RCT sont en soi un outil statistique neutre, mais comme le fait remarquer Dean Spears[6], « les RCT fournissent un discours tout prêt et de haut niveau » qui permet « une légitimation mutuelle entre les bailleurs de fonds, les chercheurs et les gouvernements ». Lorsque la méthodologie des RCT est utilisée comme un outil pour « déterminer ce qui fonctionne », sans tenir compte des libertés dans sa définition de ce qui fonctionne, alors elle risque de cautionner l'oppression.

Remerciements

Je remercie vivement Nancy Cartwright, Anne Case, Shoumitro Chatterjee, Nicolas Côté, Jean Drèze, William Easterly, Reetika Khera, Lant Pritchett, Dean Spears et Bastian Steuwer pour leurs commentaires (généreux et utiles) apportés sur une version antérieure, ainsi que pour leur réactivité.

5. Hamid et Sabah (2019) ; « Why I resigned from the Gates Foundation », *New York Times*, 26 septembre ; « Dismay at Gates Foundation prize for Narendra Modi », *The Guardian*, Lettre, 23 septembre 2019 ; « Bill and Melinda Gates Foundation under fire for award to Narendra Modi », *The Guardian*, 12 septembre 2019.

6. Dean Spears, communication personnelle, 14 octobre 2019. Avec son aimable autorisation à le citer.

Partie I

Que peuvent les RCT ?

Chapitre 1

Les *randomistas* doivent-ils (continuer à) faire la loi ?

Martin RAVALLION

Introduction

Depuis le début du nouveau millénaire, les évaluations d'impact se sont multipliées de manière exponentielle dans les pays en développement, généralement dans le but d'améliorer l'élaboration des politiques. L'International Initiative for Impact Evaluation (3ie) a compilé des métadonnées sur ces évaluations, illustrées dans la fig. 1[1] (CAMERON *et al.*, 2016 ; SABET et BROWN, 2018). Depuis l'an 2000, le nombre d'évaluations d'impact a été multiplié par 30 par rapport aux 19 années précédentes[2], une augmentation spectaculaire.

Comme le montre également la fig. 1[3], il existe deux méthodes principales. Dans la première, certaines unités se voient attribuer de manière aléatoire l'accès à un programme (le « traitement »), tandis que d'autres servent de groupe contrôle. Pour mesurer l'impact du programme, on compare ensuite les résultats moyens de ces deux échantillons. Il s'agit de la version la plus simple d'une évaluation par assignation aléatoire (*Randomized Controlled Trials* – RCT). La

1. Les chiffres couvrent la période 1981-2015.

2. La base de données de 3ie recense 4 501 évaluations d'impact sur la période 1981-2015, dont 4 338 ont été publiées entre 2000 et 2015. Depuis l'an 2000, 271 évaluations d'impact sont réalisées en moyenne tous les ans, contre 9 entre 1981 et 1999.

3. La série de 3ie est basée sur la recherche de mots-clés sélectionnés dans des textes numérisés. Le personnel de la série de 3ie m'a averti (dans une correspondance) que leurs anciens protocoles de recherche étaient probablement moins efficaces pour récupérer les études non expérimentales que les RCT antérieurs à 2000. Par conséquent, les chiffres des évaluations non randomisées de la fig. 1, qui sont plus anciens et plus bas, peuvent être trompeurs. Les comptages de 3ie recensent beaucoup plus de RCT que celles rapportées dans BOUGUEN *et al.* (2019) (ces derniers donnant les totaux cumulés et non les flux annuels).

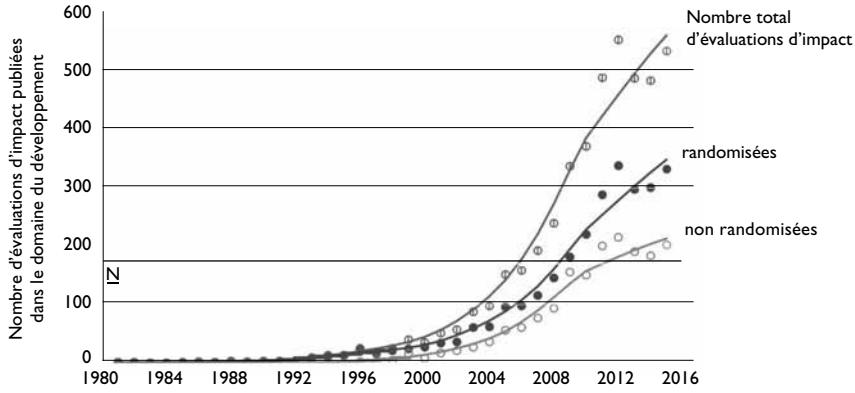


Figure 1

Nombre d'évaluations d'impact publiées chaque année pour les pays en développement.

Source : Martin Ravallion, sur la base des données de la 3ie.

Note : les courbes ajustées correspondent aux nuages des points voisins les plus proches. Voir la note 3 de ce chapitre sur la possible sous-estimation des évaluations non randomisées au cours des années précédentes.

deuxième méthode n'a pas recours à la randomisation. On y trouve notamment des études purement non expérimentales (*observationnal*) : l'attribution du traitement est considérée comme une donnée, elle est intentionnelle et non pas aléatoire. Le second groupe comprend également des attributions déterministes, par exemple basées sur des *a priori* concernant les bénéfices potentiels du traitement. Alors que certains essais non randomisés qui servent à orienter l'élaboration des politiques sont purement descriptifs, d'autres cherchent à contrôler les différences avant traitement entre les unités traitées et non traitées en se basant sur ce qui peut être observé dans les données, dans le but de tirer des conclusions causales crédibles sur l'impact.

Si les RCT ont commencé à être utilisées dans le domaine du développement vers 1980, ce n'est qu'une vingtaine d'années plus tard qu'elles se sont vraiment intensifiées. Depuis 2000, environ 60 % des évaluations d'impact reposent sur la randomisation. Le dernier recensement de 3ie compte 333 articles ayant recours à cet outil pour l'année 2015[4]. Le taux de croissance est saisissant. En comparant sur la fig. 1 les chiffres des RCT à une tendance exponentielle (et l'ajustement est bon), on constate un taux de croissance annuel d'environ 20 %, soit plus du double du taux de croissance de l'ensemble des publications scientifiques depuis la Seconde Guerre mondiale[5]. À titre indicatif, si l'on tape

4. Pour donner une idée aux économistes, cela correspond à peu près au nombre total d'articles (dans tous les domaines) publiés chaque année dans l'*American Economic Review*, le *Journal of Political Economy*, le *Quarterly Journal of Economics*, *Econometrica* et la *Review of Economic Studies* (Card et DellaVigna, 2013).

5. La régression logarithmique du nombre de RCT dans le temps donne un coefficient de 0,20 (s.e. = 0,01 ; n = 32 ; R^2 = 0,96) ou de 0,18 (0,01 ; n = 16 ; R^2 = 0,96) si on ne prend la série qu'à partir de 2000. À l'époque moderne (depuis la Seconde Guerre mondiale), le taux de croissance des publications scientifiques est estimé à 8-9 % par an (Bornmann et Mutz, 2014).

« RCT » ou « *randomized controlled trials* » dans Google Ngram Viewer, on constate que la fréquence de ce groupe de mots (par rapport à l'ensemble des ngrams dans les textes numérisés) a tendance à augmenter dans le temps et se révèle plus élevée à la fin de la série chronologique disponible (2008) qu'elle ne l'a jamais été auparavant.

Avant 2000, on n'aurait certainement pas pu prévoir que les RCT allaient connaître une croissance telle que celle illustrée par la fig. 1. En effet, pour bon nombre des opérations réalisées par les gouvernements et autres acteurs dans le domaine du développement, les RCT ne sont pas adaptées. Les RCT n'ont pas toujours été bien accueillies non plus. Il leur est souvent reproché de priver d'un programme, à des fins de recherche, certaines personnes qui en ont besoin au profit d'autres qui n'en ont pas besoin. Il fut un temps où les RCT sur le développement étaient difficiles à vendre. Mais quelque chose a changé. Comment les RCT sont-elles devenues si populaires ? Leur popularité est-elle justifiée ?

Les partisans des RCT ont été surnommés les « *randomistas*[6] ». Ils considèrent les RCT comme étant « l'étalon-or » des évaluations d'impact – l'approche la plus « scientifique » et « rigoureuse », garantissant une évaluation d'impact essentiellement athéorique et sans hypothèse, mais fiable[7]. Ce credo, émanant d'éminents économistes universitaires, a ensuite gagné le discours populaire, avec une influence notable auprès des médias, des agences de développement et des bailleurs de fonds, ainsi que parmi les chercheurs et leurs employeurs[8]. Ils privilégient les RCT de manière inconditionnelle. Si le contexte dans lequel s'inscrit une évaluation d'impact peut être très varié (types d'interventions, secteurs de l'économie, pays, communautés, groupes sociaux/ethniques), ce statut d'étalon-or est généralement revendiqué, quel que soit le contexte.

Il y a eu quelques résistances. Les RCT menées dans le domaine de la politique sociale ont soulevé de nombreuses interrogations (Heckman et Smith, 1995 ; Grossman et Mackenzie, 2005 ; Cartwright, 2007 ; Ravallion, 2009a ; 2009b ; 2012 ; Rodrik, 2009 ; Barrett et Carter, 2010 ; Deaton, 2010a ; Keane, 2010 ; Baele, 2013 ; Basu, 2014 ; Mulligan, 2014 ; Pritchett et Sandefur,

6. Ce terme « *randomistas* » n'est pas péjoratif. Les partisans des RCT l'emploient également volontiers, comme Leigh (2018).

7. Par exemple, Banerjee (2006) écrit que : « Les expérimentations randomisées comme celles-ci – c'est-à-dire les essais où l'intervention est attribuée au hasard – constituent le meilleur moyen et le plus simple d'évaluer l'impact d'un programme. » De même, Imbens (2010 : 407) affirme que « les expériences randomisées tiennent effectivement une place particulière dans la hiérarchie des preuves : elles se situent au sommet ». Et Duflo (2017 : 3) parle des RCT comme étant « l'outil de choix ».

8. Une phrase de la page Wikipedia en anglais sur les évaluations d'impact témoigne de l'influence accrue de l'étalon-or : « Les expérimentations randomisées sur le terrain sont les plans de recherche les plus performants pour évaluer l'impact d'un programme […] car elles permettent une estimation juste et précise des effets réels d'un programme. » Dans un autre exemple, Keating (2014) écrit que « les *randomistas*, partisans des expérimentations randomisées, ont récemment transformé notre conception du développement économique et de l'aide aux pays pauvres ». Dans le même ordre d'idées, l'ouvrage de Leigh (2018) s'intitule *Randomistas : comment les chercheurs radicaux ont changé notre monde*.

2015 ; Favereau, 2016 ; Ziliak et Teather-Posadas, 2016 ; Hammer, 2017 ; Deaton et Cartwright, 2018 ; Gibson, 2019 ; Young, 2019 ; voir aussi Pritchett, chap. 2, ce volume). Certains détracteurs ont notamment fait valoir que, en premier lieu, les hypothèses requises pour une estimation fiable de l'impact par le biais d'une RCT ne sont pas forcément fondées dans la réalité ; en second lieu, les RCT sont contestables sur le plan éthique ; et enfin, le caractère « boîte noire » des RCT limite leur utilité pour l'élaboration de politiques, y compris pour ce qui est de leur transposition à plus grande échelle et de la difficulté à déterminer le potentiel d'impact dans d'autres contextes. Ces critiques ont suscité des réactions de la part de certains défenseurs des RCT (Banerjee et Duflo, 2009 ; Imbens, 2010 ; 2018 ; Goldberg, 2014 ; Glennerster et Powers, 2016 ; McKenzie, 2019), l'un d'entre eux ayant qualifié de « foutaises » les critiques d'ordre éthique émises à l'encontre des RCT (Fiennes, 2018), alors qu'un détracteur a même qualifié la révolution des RCT de « folie » (et même « bien pire qu'une folie ») (Pritchett, chap. 2, ce volume).

Face à l'importance croissante des RCT dans le domaine du développement et aux débats qui se poursuivent, ce chapitre revient, dix ans plus tard, sur la question posée par Ravallion (2009a), « Les *randomistas* doivent-ils faire la loi ? ». Les *randomistas* « font la loi » au sens où ils revendiquent une hiérarchie des méthodes, qui est le fondement de leur autorité intellectuelle et du pouvoir de persuasion qu'ils exercent[9]. Cette hiérarchie est l'objet principal de ce chapitre. Tout en reconnaissant l'intérêt des RCT à certaines fins, ce chapitre montre que l'engouement qui a émergé en faveur des RCT ne repose pas sur une juste appréciation des limites de cet outil de recherche. Ce chapitre n'est pas destiné aux experts des deux camps, mais, plus largement, à la communauté des économistes et aux autres spécialistes des sciences sociales, aux bailleurs de fonds, aux décideurs politiques et à leurs conseillers, aux étudiants et aux jeunes chercheurs.

J'ouvrirai ce chapitre par un aperçu de la théorie de l'évaluation d'impact, et notamment des éléments pertinents pour le choix des méthodes. J'aborderai dans le développement suivant l'influence des *randomistas* sur la recherche pour le développement, les inquiétudes quant à la validité éthique de leur méthode de prédilection, avant d'envisager la pertinence de leur recherche sur le plan politique, puis de conclure.

Fondements de l'évaluation d'impact

Il s'agit d'attribuer des programmes : certaines unités (les « traités ») d'une population bien définie bénéficient du programme et d'autres non. Imaginez que l'on prenne deux échantillons aléatoires dans la population, l'un du groupe

9. L'observation de McKenzie (2019) selon laquelle seuls 10 % de tous les articles sur l'économie du développement (tous domaines confondus, dans 14 revues) portent sur des RCT ne permet donc pas de réfuter la revendication de la suprématie des *randomistas*, dans le sens utilisé ici.

traité et l'autre du groupe non traité, puis que l'on mesure les résultats pertinents pour les deux. Il s'agit là d'une expérimentation unique[10]. La différence entre les résultats moyens correspond à l'estimation de l'impact moyen réel de l'essai pour cette population, également appelé effet moyen du traitement (*Average Treatment Effect* – ATE). Cette estimation peut différer de la valeur réelle pour diverses raisons : erreurs de mesure, variabilité de l'échantillonnage, effets d'entraînement (« contamination ») entre les deux groupes, effets de surveillance et/ou biais systématique résultant de toute variable de confusion qui modifierait simultanément les résultats et l'état du traitement. La paire d'échantillons de chaque essai donne une estimation différente, parfois trop élevée, parfois trop faible, même s'il nous est impossible de savoir de combien, puisque nous ne connaissons (évidemment) pas la valeur réelle. Tout essai comporte son lot d'erreurs.

La RCT idéale correspond à la configuration décrite ci-dessus, où le statut de traitement de l'essai est également attribué de manière aléatoire (en plus du tirage d'échantillons aléatoires dans les deux populations, l'une traitée et l'autre non) et où la seule erreur est due à la variabilité de l'échantillonnage. Cet idéal peut être illusoire en pratique, surtout avec des sujets humains. Nous reviendrons plus tard sur les écarts entre la réalité et l'idéal, mais pour l'instant, nous allons nous baser sur une RCT idéale. Dans ce cas particulier, à mesure que le nombre d'essais augmente, la moyenne de leurs estimations tend à se rapprocher de l'impact moyen réel. C'est en cela que l'on dit qu'une RCT idéale est impartiale, c'est-à-dire que l'erreur expérimentale est ramenée à zéro *en espérance.* Cette propriété nous permet également d'estimer la variance des estimations. En utilisant à la fois le traitement aléatoire et l'échantillonnage aléatoire, il est donc plus facile de calculer l'erreur type de l'estimation de l'impact issue d'une RCT, afin d'établir un intervalle de confiance statistique[11].

Les grands *randomistas* ont parfois omis le qualificatif « en espérance », ou ignoré ses implications sur l'existence d'erreurs expérimentales (comme le notent Deaton et Cartwright, 2018). Pour ces défenseurs des RCT, toute différence dans les résultats moyens entre les échantillons de traitement et de contrôle est attribuable à l'intervention[12]. Cette erreur courante doit être considérée comme

10. Le terme « expérimentation » est parfois défini comme une situation sous le contrôle total de l'évaluateur, ce qui est le cas avec une RCT ; voir, par exemple, Cox et Reid (2000). Cependant, il est extrêmement rare que l'évaluateur contrôle tout dans les RCT impliquant des sujets humains, comme c'est le cas pour les évaluations de politiques sociales. Dans ce chapitre, j'utilise le terme « expérimentation » dans son sens le plus large, en excluant cette idée de contrôle total. Il peut s'agir ou non d'une RCT.

11. Les pratiques actuelles à cet égard peuvent être remises en question. Young (2019) soulève un certain nombre de préoccupations d'erreurs types dans des estimations d'impact faites dans le passé au moyen de RCT avec des contrôles de régression et montre que de nombreuses publications économiques ont surestimé la signification statistique de leurs estimations d'impact. Voir également les discussions dans Deaton et Cartwright (2018) et Imbens (2018).

12. Par exemple, en ce qui concerne les RCT, Banerjee et Duflo (2017) affirment que « toute différence entre le groupe de traitement et le groupe de contrôle peut être attribuée avec certitude au traitement ». Banerjee *et al.* (2019a) soutiennent quant à eux qu'une RCT garantit que « toute différence entre les unités de traitement et de contrôle reflète l'impact du traitement ». [suite p. suiv.]

plus qu'une simplification mineure à fin d'exposition[13]. Cependant, cette simplification est désormais largement ancrée dans le discours public. Outre les experts (si l'on met de côté leurs déclarations non vérifiées), de nombreuses personnes dans la communauté du développement pensent désormais que toute différence mesurée dans une RCT entre le groupe de traitement et le groupe de contrôle est imputable au traitement. Pourtant, ce n'est pas le cas ; même la RCT idéale comporte une erreur expérimentale inconnue.

Le cas sans traitement est rare, mais instructif. En l'absence de tout autre effet lié à l'attribution (tel que le suivi), l'impact est nul. Pourtant, l'erreur aléatoire d'un essai peut toujours donner un impact moyen non nul lors d'une RCT. Citons par exemple une RCT menée au Danemark, dans laquelle 860 personnes âgées ont été réparties au hasard et sans le savoir dans des groupes de traitement et de contrôle pendant une période de 18 mois sans aucune intervention réelle (Vass, 2010). À l'issue de cette période, on a observé une différence statistiquement significative (prob. = 0,003) des taux de mortalité.

À la lumière de ces observations, il convient de réfléchir au choix des méthodes. Supposons que, avec un budget donné, nous puissions mettre en œuvre soit une RCT, soit une étude non expérimentale. Dans le cas de l'étude non expérimentale, les sujets sont sélectionnés pour le programme. Prenons des échantillons aléatoires de ceux qui participent et de ceux qui ne participent pas. Nous voulons classer les méthodes *ex ante* selon le degré auquel les estimations de l'essai sont susceptibles de se rapprocher de la valeur réelle. Disons qu'une estimation est « proche de la vérité » si elle se situe dans un intervalle fixe centré sur la valeur réelle. L'accent est mis ici sur la « validité interne » de chaque estimateur – sa précision pour la population en question. La « validité externe » est traitée *infra*.

Si ce statut d'étalon-or est si souvent invoqué, c'est en raison de l'absence de biais dans la RCT idéale. Les économistes se sont beaucoup concentrés sur une source particulière de biais, à savoir la différence entre l'espérance mathématique de l'estimation d'un paramètre et sa valeur réelle (inconnue). Dans certains ouvrages, on appelle cela le « biais systématique », par opposition aux sources d'erreurs spécifiques aux essais, potentiellement nombreuses[14]. Même selon cette définition étroite, une étude non expérimentale n'est pas nécessairement biaisée. On ajuste généralement le déséquilibre des covariables, y compris dans une RCT. Le biais est supprimé lorsque le statut du traitement est conditionnellement exogène, c'est-à-dire non corrélé avec le terme d'erreur conditionnel aux covariables (bien que cette hypothèse soit clairement plus solide que pour

On trouve une affirmation semblable, mais sans fondement, dans « l'Introduction à l'évaluation » du site web du Abdul Latif Jameel Poverty Action Lab (J-PAL) (sur laquelle nous revenons *supra*). Après avoir présenté les grandes lignes d'une RCT réalisée dans le cadre d'un projet de purification de l'eau, avec des groupes de traitement et de contrôle, le J-PAL précise que « toute différence constatée par la suite peut être attribuée au fait que l'un a bénéficié du programme de purification de l'eau, et l'autre non ». On trouve un autre exemple dans un manuel technique sur l'évaluation d'impact de la Banque interaméricaine de développement et de la Banque mondiale (Gertler *et al.*, 2016).

13. Comme le suggère Imbens (2018), dans ses commentaires sur Deaton et Cartwright (2018).

14. Sur les multiples sources de « biais », voir Hernánet Robins (2018).

une RCT). Cette hypothèse peut être acceptable ou non, selon le contexte (le programme et les données disponibles). Le caractère exogène ou non du traitement, compte tenu des variables de contrôle, dépend de la capacité de celles-ci à refléter correctement les facteurs déterminants du placement en traitement ; il convient d'en juger pour chaque contexte. Pour déterminer les données dont on a besoin, il faut bien comprendre les déterminants économiques et sociaux du placement dans le cadre du programme, c'est-à-dire les problèmes de décision auxquels sont confrontés les différents acteurs dans le contexte spécifique. Souvent, des facteurs de confusion qui ont été omis subsisteront, même si cela n'implique pas nécessairement des biais importants en ajustant seulement pour les facteurs de confusion observés.

Si les facteurs de confusion non mesurés sont très préoccupants, il est possible d'éliminer le biais en trouvant une source de variation exogène dans le statut du traitement qui ne soit pas non plus un déterminant des résultats du traitement donné. Il s'agit d'une variable instrumentale (VI). Une VI valide doit être corrélée avec le statut du traitement *et* non corrélée avec les résultats, le traitement donné et les variables de contrôle. Dans une régression, cela exige que la VI ne soit pas corrélée avec le terme d'erreur – donnant ce qu'on appelle la « restriction d'exclusion ». Cette condition doit finalement être jugée sur des bases théoriques, même s'il peut être utile d'étudier attentivement les facteurs déterminant le statut de traitement dans le cadre spécifique pour trouver des VI théoriquement plausibles, ainsi que des variables de confusion potentielles. Prenons par exemple un programme dont l'attribution du traitement dépend du fait que le score d'éligibilité soit supérieur à un seuil critique, ainsi que d'autres facteurs rendant l'attribution plus floue. Tant que le seuil est arbitraire (c'est-à-dire que les résultats contrefactuels moyens ne changent pas pour ce qui est du seuil), le fait que le score soit supérieur ou inférieur à cette valeur critique constitue une VI plausible en théorie[15]. Bien que moins connu des économistes, le biais lié à des facteurs de confusion non mesurés dans une étude non expérimentale peut être éliminé s'il existe une variable intermédiaire qui relie le traitement aux résultats, mais ne dépend pas des facteurs de confusion[16].

Même si nous sommes d'accord sur le fait qu'une RCT permet d'éliminer plus efficacement les biais dans un contexte spécifique, cela ne justifie pas pour autant sa suprématie, et ce, pour deux raisons majeures. Premièrement, étant donné les contraintes auxquelles sont soumises les RCT dans la pratique, il peut s'avérer impossible de représenter correctement la population concernée. Au moins, lorsque les médias sont libres, les gouvernements sont conscients des risques politiques encourus s'ils soutiennent une recherche douteuse sur le plan éthique. Si les RCT sont parfois réalisées de concert avec les gouvernements, il est souvent plus facile

15. Il s'agit d'un exemple de régression par discontinuité ; pour un traitement formel, voir Hahn *et al.* (2001).

16. C'est ce que l'on appelle parfois le réglage « *front-door* » par opposition au réglage « *back-door* », qui utilise une VI (Pearl et Mackenzie, 2018) (voir les chap. 4 et 7, ce volume). Pour un exemple de réglage « *front-door* » voir Glynn et Kashin (2018). Pour un traitement plus formel, voir Pearl (2009 : chap. 3).

d'accepter des études non expérimentales plus bénignes. C'est pourquoi les *randomistas* universitaires à la recherche de partenaires locaux jugent plus intéressant de travailler avec des organisations non gouvernementales (ONG) locales. La randomisation peut ainsi permettre de fournir (dans des conditions idéales) une estimation d'impact non biaisée pour une sous-population sélectionnée de manière non aléatoire, comme celle qui se trouve dans la région d'une ONG locale coopérative. En outre, il arrive que le processus de sélection du sous-échantillon compatible soit loin d'être clair (et, souvent, l'article académique correspondant ne mentionne même pas la manière dont il a été choisi). Il est difficile de savoir ce que l'on peut tirer d'une estimation non biaisée pour un sous-échantillon de la population sélectionné de manière non aléatoire. Les impacts potentiels sont si hétérogènes qu'une étude non expérimentale biaisée d'un échantillon aléatoire de toute la population pourrait s'avérer plus proche de la vérité.

Deuxièmement, le biais n'est pas la seule chose qui compte. La règle de décision à adopter pour choisir un estimateur (et plus généralement pour concevoir une étude) dépend de l'application. Une règle de décision statistique courante consiste à minimiser l'erreur quadratique moyenne (EQM), c'est-à-dire la valeur attendue de l'écart quadratique entre l'estimation et sa valeur réelle. Comme on le sait en statistique, l'EQM est le biais de l'estimateur au carré *plus* sa variance[17]. Cette règle de décision ne nous dit donc pas qu'un estimateur non biaisé est toujours le meilleur[18]. L'EQM n'est pas la seule règle de décision défendable – par exemple, on pourrait se demander à quelle fréquence les essais se situent dans une certaine distance absolue de la valeur réelle. Ce que je cherche à démontrer, c'est que l'absence de biais ne suffit pas.

C'est ici que la dimension économique de l'évaluation d'impact entre en jeu. La variance des estimations diminue lorsque la taille des échantillons est plus importante. De nombreuses études non expérimentales utilisent des données existantes, provenant de registres administratifs (*big-data*), ainsi que des enquêtes existantes. Les RCT nécessitent généralement de nouvelles enquêtes spécifiques. Ainsi, pour un budget donné, les RCT auront souvent des échantillons de taille plus faible et (donc) des variances plus élevées.

Le résultat n'est pas non plus évident lorsqu'un essai non randomisé nécessite de nouvelles enquêtes. Pour réduire les biais, il faut disposer de meilleures données. Avec des questionnaires d'enquête plus longs, la taille des échantillons sera probablement réduite pour un budget donné. Mais il est peu probable que les exigences en matière de données pour une RCT soient différentes. Il faut en effet disposer de données de base pour tester l'équilibre des covariables d'une

17. Si β est la vraie valeur et $\hat{\beta}$ son estimateur, alors (par définition) :
$EQM = E((\hat{\beta} - \beta)^2) = (E(\beta) - \beta)^2 + E((\hat{\beta} - E(\hat{\beta}))^2)$. Le premier terme du membre droit correspond au biais de β au carré, tandis que le second terme correspond à sa variance.

18. Les textes d'introduction à l'économétrie le montrent bien. Par exemple, selon JONSTON (1984 : 28) « en ce qui concerne le critère de l'erreur quadratique moyenne, il peut être préférable d'avoir un estimateur biaisé plutôt qu'un estimateur avec un biais plus faible ou nul si sa variance est suffisamment faible pour compenser le biais plus important ». Voir également la discussion à ce sujet dans GREEN (1991 : 97-99).

RCT[19]. La randomisation supplémentaire (d'un traitement) dans le cadre d'une RCT implique très souvent des frais, et il pourrait bien s'avérer nécessaire d'effectuer une nouvelle randomisation pour assurer l'équilibre des covariables (Morgan et Rubin, 2012). Dans le domaine médical, on considère généralement que les RCT sont plus coûteuses que les études non expérimentales (Hannan, 2008 ; Frieden, 2017). Je ne dispose pas de données précises sur les coûts des évaluations d'impact dans le domaine du développement, mais on entend souvent des voix s'élever pour dénoncer la faiblesse statistique des RCT[20]. Si l'on compare (sommairement) les coûts des évaluations de la Banque mondiale, on constate que les RCT coûtent plus cher[21]. Les RCT dans le domaine du développement rencontrent souvent des difficultés de mise en œuvre sur le terrain que l'on ne retrouve pas dans les études non expérimentales. Pour mieux comprendre ce que cela peut signifier pour le choix de la méthode, supposons que chaque essai soit réalisé selon l'une des deux méthodes de distribution normale, à savoir l'une pour une expérimentation randomisée et l'autre pour une expérimentation non randomisée. Les paramètres (moyenne et variance) dépendent de la méthode choisie. La moyenne de la distribution des résultats d'une RCT est considérée comme la véritable moyenne, alors qu'elle ne l'est pas pour les essais non randomisés. Malgré le biais, la variance d'un essai non randomisé pourrait être suffisamment faible pour garantir qu'il soit plus proche de la vérité qu'un essai randomisé. La fig. 2 illustre un cas hypothétique et montre que même une étude non expérimentale biaisée peut être plus proche de la vérité qu'une RCT (non biaisée)[22]. Les courbes de densité illustrent les estimations d'impact des plans RCT et non RCT, toutes deux fondées sur des distributions normales (les densités peuvent ou non être conditionnées par des covariables). L'impact réel est nul, ce qui correspond à la moyenne de la distribution sur laquelle sont établies les RCT. Les essais non randomisés sont plutôt basés sur une distribution dont la moyenne est de -0,5, soit leur biais systématique. Autre différence : les RCT sont basées sur une distribution avec une variance de 2, alors que, pour l'étude d'observation, elle est de 1. Autrement dit, pour un budget donné, les essais non randomisés permettent de doubler la taille de l'échantillon pour chaque essai.

19. Pour les RCT, il est généralement recommandé de réaliser des tests d'équilibrage *ex post*, ainsi que des ajustements rétrospectifs (Cox et Reid, 2000 ; Hinkelmann et Kempthorne, 2008 ; Bruhn et McKenzie, 2009 ; Hernán et Robins, 2018).

20. Par exemple, au sujet des RCT dans le domaine du développement, White (2014) indique que « la puissance statistique réelle de nombreuses RCT n'est que de 50 % environ. Ainsi, le recours à une RCT ne vaut pas mieux que de jouer à pile ou face pour déterminer si une intervention fonctionne ou pas ». La variabilité de l'échantillonnage semble expliquer la moitié ou plus de la variabilité des estimations d'impact des RCT ; voir Meager (2019) en référence aux programmes de microcrédit.

21. Ces derniers temps, les évaluations d'impact de la Banque mondiale étaient généralement des RCT dont le coût moyen était considérablement plus élevé que les évaluations réalisées dans le cadre de la coopération financière internationale (au sein du groupe de la Banque mondiale), où les études non expérimentales sont plus courantes (World Bank, 2012). Ceci n'est précisé qu'à titre indicatif, puisque la comparaison n'est pas dûment contrôlée.

22. Green (1991, section 4.0.3) utilise un exemple similaire pour montrer qu'un estimateur plus biaisé peut avoir une EQM plus faible.

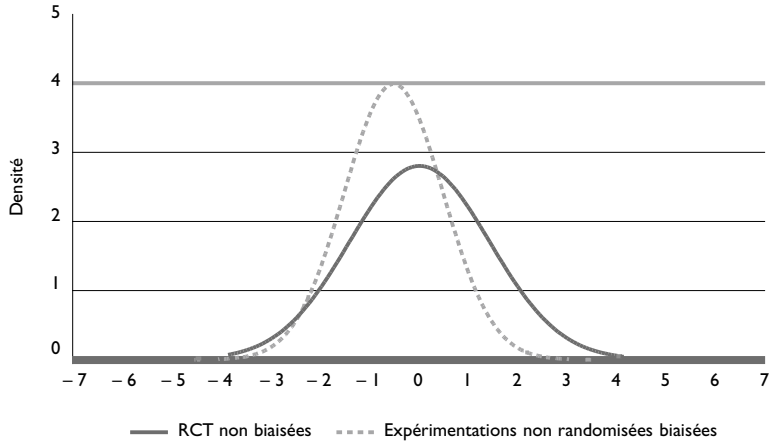


Figure 2
Fonctions de densité pour les estimations de l'impact moyen selon deux modèles hypothétiques d'évaluations d'impact.
Source : Martin Ravallion.

Quelle méthode est donc la plus efficace en fournissant des estimations qui tendent à être plus proches de la vérité ? On entend par « plus proche de la vérité » la probabilité de se trouver dans un intervalle fixe centré sur la valeur réelle – dans ce cas, un intervalle $(-\delta, \delta)$ (où $\delta > 0$)[23]. La fig. 3 donne le pourcentage d'essais proches de la vérité pour chaque méthode. Supposons que nous définissions « plus proche de la vérité » comme une estimation de l'impact dans l'intervalle (– 0,5, 0,5). Lorsque la méthode randomisée est utilisée, nous constatons que, dans 27 % des essais, l'estimation se situe dans cet intervalle, contre 34 % pour la méthode non randomisée. En revanche, si nous définissons « plus proche de la vérité » comme une estimation comprise dans l'intervalle (– 1, 1), alors c'est le cas pour 52 % des RCT contre 62 % pour les études observationnelles. Dans cet exemple, on constate que l'étude non expérimentale est plus proche de la vérité, quelle que soit la δ !

Bien sûr, ce n'est là qu'une des nombreuses possibilités, et on peut facilement construire des exemples où les RCT sont plus concluantes. Les fig. 2 et 3 montrent seulement qu'une évaluation d'impact moins biaisée ne nous rapproche pas nécessairement de la vérité. Cette question demeure ouverte, car elle dépend essentiellement du degré de performance des essais que l'on peut se permettre selon le budget. Ce qui est important à retenir ici, c'est que, pour un budget donné, les RCT peuvent tout à fait aboutir à des estimations d'impact moins fiables, qui sont souvent plus éloignées de la vérité que celles d'une étude non expérimentale, même biaisée. Ainsi, le statut « d'étalon-or » revendiqué (de manière inconditionnelle) ne repose sur aucune justification théorique.

23. Pour certaines applications, l'intervalle n'a pas besoin d'être symétrique par rapport à la valeur réelle, c'est-à-dire que les erreurs dans un sens sont plus coûteuses par rapport à l'objectif convenu.

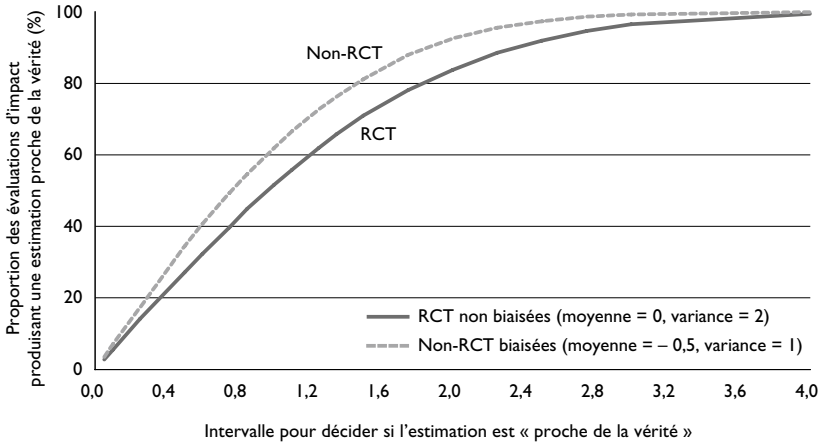


Figure 3

Proportion d'essais donnant une estimation de l'impact proche de la vérité, avec une comparaison entre une RCT non biaisée et une expérimentation non randomisée biaisée sur un échantillon plus large.

Source : Martin Ravallion.

Pour la défense des RCT, nous disposons actuellement de peu d'informations sur la distribution des biais des études non expérimentales, alors que, comme nous l'avons noté, l'absence de biais d'une RCT est assortie d'une variance estimable, ce qui facilite le calcul de son intervalle de confiance. Cela souligne la nécessité d'effectuer davantage de recherches sur la distribution des estimations issues des études non expérimentales, par exemple en comparant les estimations avec celles des RCT pour un même contexte (voir par exemple Chabé-Ferret, 2018). Cela dit, si nous continuons à ne réaliser que des RCT (lorsque c'est possible), nous risquons de négliger les études non expérimentales, lesquelles s'avèrent plus souvent proches de la vérité.

Si l'on connaît suffisamment bien le contexte et le programme pour identifier les facteurs de confusion pertinents – le modèle de fonctionnement du programme – et si l'on peut recueillir des données à leur sujet ou trouver des facteurs mesurables pour éliminer la confusion, il est possible de mieux appréhender la situation et obtenir ainsi une estimation très fiable de l'impact par simple observation. D'un autre côté, si les possibilités de collecter des données de base sur les facteurs de confusion pertinents sont limitées et si le coût unitaire de la randomisation n'est pas trop élevé (de sorte qu'il est possible d'obtenir des échantillons de taille raisonnable avec le budget disponible), alors il peut être très intéressant de réaliser une RCT. Ce qui est remis en question ici, c'est la fameuse généralisation de « l'étalon-or ».

Nous pouvons aller plus loin et nous demander quelle est la méthode de conception optimale pour minimiser (par exemple) l'EQM, sans pour autant dissimuler

notre incertitude quant au véritable modèle[24]. Admettons que quelques-unes au moins des données de base soient des covariables continues et que nous ayons des *a priori* bayésiens sur l'incertitude du modèle. Nous pouvons alors nous appuyer sur un résultat de KASY (2016), à savoir qu'il existe une attribution déterministe (non aléatoire) du statut de traitement basé sur les covariables qui permet de minimiser l'EQM attendue[25]. Il n'y a donc aucun avantage à randomiser l'attribution en fonction des covariables (alors que la prise en compte des éléments observables dans les RCT entraînera des gains d'efficacité). Par conséquent, pour justifier une nette préférence aux RCT, il faut accorder à la randomisation une certaine valeur intrinsèque en tant que fin en soi, et être prêt à renoncer à la précision en matière du véritable impact moyen. Les partisans de cette méthode se sont parfois retranchés sur ces préférences méthodologiques, sans tenir compte de la précision de l'estimation de l'impact ; par exemple, BANERJEE *et al.* (2017b) montrent que les RCT peuvent continuer à dominer tant que l'on accorde un poids suffisamment important au bien-être de ceux qui préfèrent les RCT.

L'influence des *randomistas* provient en partie de la croyance (très répandue) selon laquelle les RCT, lorsqu'elles s'y prêtent, constituent le meilleur outil statistique. Cette analyse de la théorie sous-jacente a jeté un doute considérable sur cette croyance. Comme nous le verrons, les autres sources de leur influence sont tout aussi discutables.

L'influence des *randomistas* sur la recherche pour le développement

Parmi les premiers exemples de RCT menées dans le cadre de politiques sociales, on peut citer les diverses expérimentations réalisées en matière de politiques sociales aux États-Unis à partir des années 1960[26]. En ce qui concerne les applications plus récentes dans le domaine du développement, la base de données de 3ie compte 133 RCT publiées au cours de la période 1981-1999. La toute première RCT enregistrée dans la base de données concerne un projet de recherche de la Banque mondiale sur les interventions en matière d'éducation

24. Selon KASY (2016), cela peut être considéré comme un problème dans la théorie de la décision statistique, c'est-à-dire le fait de choisir une méthode d'estimation pour minimiser une fonction de perte à partir des données réellement disponibles.

25. Cela vaut pour toute fonction de risque bayésienne et pour la règle de décision « minimax » dans le pire des cas (KASY, 2016). Kasy fournit un logiciel permettant de mettre en œuvre l'attribution optimale du traitement afin de minimiser l'EQM. Notez qu'une covariable continue garantit une attribution optimale unique du statut de traitement. Avec des covariables discrètes, une RCT peut s'avérer tout aussi performante, mais pas plus performante.

26. Sur les RCT menées par le passé dans le cadre de la politique sociale américaine, voir les discussions de BURTLES (1995) et de LIST et RASUL (2011). Pour plus de détails sur l'histoire plus générale des RCT, voir ZILIAK (2014) et LEIGH (2018).

(manuels scolaires et cours diffusés à la radio) visant à améliorer les résultats en mathématiques des élèves au Nicaragua (JAMISON *et al.*, 1981)[27]. Parmi les RCT antérieures à 2000, celle réalisée par le gouvernement mexicain pour l'évaluation de *Progresa*, qui a débuté en 1997, est un exemple particulièrement remarquable. Les résultats (généralement positifs) présentés dans la littérature et découlant des données de cette RCT ont eu une influence sur l'expansion des transferts monétaires conditionnels dans plus de 50 pays aujourd'hui[28].

Au début du nouveau millénaire, l'idée des RCT dans le domaine du développement n'était donc pas nouvelle. En revanche, elle a beaucoup gagné en popularité. La production annuelle de RCT est bien plus élevée depuis 2000 (fig. 1). De nombreux universitaires et groupes ont apporté leur contribution, le Abdul Latif Jameel Poverty Action Lab (J-PAL) étant celui qui se démarque le plus[29]. Il a été fondé en 2003, sous le nom de Poverty Action Lab, au sein du département d'économie du Massachusetts Institute of Technology (MIT) par Abhijit Banerjee, Esther Duflo, et Sendhil Mullainathan. Au moment de rédiger le présent ouvrage, le site web du J-PAL[30] comptabilise 1 012 RCT terminées ou en cours dans 83 pays. Pour un groupe de recherche universitaire, il est absolument remarquable de parvenir à un tel résultat en seulement 15 ans. En plus de ses propres RCT, le J-PAL a clairement influencé le changement d'orientation de l'économie empirique du développement vers cette méthode d'expérimentation. En effet, malgré ses innombrables expérimentations randomisées, il est peu probable que le J-PAL représente encore la majorité des RCT en cours aujourd'hui.

Ces efforts (centrés sur le J-PAL, mais qui vont encore plus loin) ont apporté plus de prestige à ce type de recherche empirique sur l'économie du développement, comme en témoigne l'attribution du prix 2019 de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel à Abhijit Banerjee, Esther Duflo et Michael Kremer. Comme l'indique le titre du communiqué, le prix a été décerné « pour leur approche expérimentale de la lutte contre la pauvreté ».

Dans cette section, nous examinerons dans un premier temps les raisons de l'influence des *randomistas* sur la recherche pour le développement et nous nous demanderons ensuite si elle est justifiée.

27. BANERJEE *et al.* (2019a) affirment que l'utilisation des RCT dans le domaine du développement a été « déclenchée » en 1994 par une étude de l'un des auteurs (Kremer). Mais cela faisait au moins 13 ans que des RCT dans le domaine du développement étaient déjà publiées.

28. Concernant l'évaluation d'impact de *Progresa* et son influence, voir SKOUFIAS et PARKER (2001) et FISZBEIN et SCHADY (2010).

29. L'organisation à but non lucratif Innovations for Poverty Action (IPA), fondée en 2002 par Dean Karlan à Yale constitue un autre groupe important qui réalise et promeut les RCT. L'IPA et le J-PAL travaillent souvent ensemble, et entretiennent des liens étroits. Au sein des organisations internationales, le groupe le plus important qui réalise des RCT est le groupe Development Impact Evaluation (DIME) de la Banque mondiale ; les trois quarts des évaluations du DIME utilisent cette méthode (World Bank, 2016).

30. https://www.povertyactionlab.org/

Pourquoi les *randomistas* ont-ils autant d'influence ?

L'opinion selon laquelle les RCT (lorsqu'elles s'y prêtent) parviennent nettement mieux à inférer la causalité que les études non expérimentales issues d'enquêtes s'est imposée assez largement. La page d'accueil du site du J-PAL indique que : « Notre mission est de réduire la pauvreté en veillant à ce que les politiques soient fondées sur des preuves scientifiques ». Pour atteindre cet objectif, le J-PAL ne réalise que des RCT. Cela ne veut pas dire que les chercheurs du J-PAL considèrent les études non expérimentales comme non scientifiques (et, indépendamment du J-PAL, de nombreux chercheurs affiliés au J-PAL utilisent des méthodes basées sur des données non expérimentales). Toutefois, dans ce contexte, les expressions « preuves scientifiques » et « preuves rigoureuses » – autre favorite, figurant notamment sur le site web du J-PAL – sont des formules toutes faites pour désigner les RCT aux yeux de la plupart des lecteurs, et cela est tout à fait intentionnel. Les implications vont bien au-delà de la revendication du statut « d'étalon-or » : pour certains de leurs défenseurs, les RCT ne sont pas seulement en tête de liste des méthodes approuvées, pour eux, il n'y a rien d'autre sur la liste !

L'attrait pour les RCT reflète en partie les défis rencontrés dans l'identification de l'impact causal. Depuis les années 1990, les problèmes d'identification en économie[31] font l'objet d'une attention accrue, dont on peut se réjouir, même si certains considèrent que cela a partiellement occulté d'autres questions importantes, telles que les erreurs de mesure (Gibson, 2019). La validité des estimateurs des VI a, quant à elle, suscité une attention plus critique. Il est facile de montrer que si l'une des conditions susmentionnées pour une VI valide n'est pas remplie, l'estimation peut être biaisée – probablement davantage que dans le cas des moindres carrés ordinaires (MCO), qui considèrent le placement comme exogène. Les chercheurs n'ont pas eu de mal à trouver des variables exogènes corrélées avec le statut de traitement choisi (même s'ils ont dû faire les tests appropriés). Il a souvent été beaucoup plus difficile d'accepter, sur le plan théorique, la restriction d'exclusion (selon laquelle la VI ne doit pas être pertinente pour les résultats, compte tenu du statut du traitement et des variables de contrôle). La VI était parfois facilement acceptée, mais ce n'était pas toujours le cas. Vers le milieu des années 1990, les participants aux conférences et les experts ont commencé à souligner régulièrement les raisons pour lesquelles la VI figurant dans certains articles spécifiques pouvait avoir un effet sur les résultats qui était indépendant de la variable endogène. Certains économistes ont ensuite commencé à rejeter presque toute tentative visant à établir une causalité sans répartition aléatoire.

Si l'on veut seulement connaître la différence des résultats moyens entre ceux qui bénéficient du traitement et ceux qui n'en bénéficient pas – ce que l'on appelle le paramètre de l'intention de traiter (*Intention To Treat* – ITT) –, alors

31. Dans l'usage courant, on dit qu'un paramètre est « identifié » lorsque sa valeur peut, en principe, être déduite des données à partir de certaines hypothèses.

la randomisation permet d'éviter ces préoccupations d'estimation de la VI. Dans le cas de la randomisation, l'attribution du traitement est exogène, sans corrélation avec le terme d'erreur de régression. Cependant, l'ITT peut être un paramètre d'intérêt assez limité. Elle est parfois présentée comme étant « pertinente pour la politique », dans la mesure où la politique correspond souvent à l'attribution de la *possibilité* de traiter. Mais comment réagiriez-vous si vous constatiez que l'impact moyen est (par exemple) nul chez les sujets à qui l'on propose un traitement, mais positif chez ceux qui le suivent ? Ce type de résultat intéresserait sûrement les décideurs politiques et les citoyens. Pour tirer parti d'une RCT, un potentiel participant au traitement voudra connaître l'impact moyen sur les personnes traitées.

L'adoption (*take up*) d'un traitement attribué de manière aléatoire à des sujets humains n'est jamais garantie et il est généralement endogène. Ainsi, le problème économétrique se pose de nouveau dans la pratique. Mais les *randomistas* ont une solution : utiliser l'attribution randomisée comme VI du traitement réel. Le *take up* est clairement subordonné à l'attribution, de sorte que cette VI est corrélée au statut du traitement. Comme elle est aléatoire, la VI n'est pas non plus corrélée (en espérance) avec le terme d'erreur lorsque l'effet du traitement est identique pour toute la population. Nous reviendrons plus tard sur les complications qui peuvent survenir lorsque les impacts varient d'une manière que le chercheur ignore, mais qui est connue de chaque participant, qui réagit en conséquence.

Au-delà de ces arguments économétriques, un certain nombre d'autres facteurs ont contribué au rayonnement croissant des *randomistas* dès le début des années 2000. Les chercheurs qui n'utilisaient pas la randomisation ont commencé à être critiqués par les *randomistas*, et leurs articles ont peu à peu cessé d'être cités dans la littérature spécialisée. Ce processus s'est déroulé en partie à travers les commentaires des relecteurs évaluant les articles pour les revues scientifiques, qui ne sont pas publics. Si les rédacteurs en chef des revues ne sont pas obligés d'accepter ce genre de critiques, les grands *randomistas* semblent avoir eu de l'influence et ont fini par occuper une place importante auprès des rédacteurs en chef et des comités de rédaction des revues économiques. Parfois, les critiques ont aussi pris une dimension publique, comme c'est le cas de l'étude de FINKELSTEIN et TAUBMAN (2015), qui remettait en question le fait que les méthodes utilisant des données d'enquêtes et d'autres méthodes non randomisées soient souvent utilisées dans l'évaluation des politiques de prestation de soins de santé. Cette conclusion a ensuite été rapportée dans le *New York Times* sous le titre « Few Health System Studies use *Top* Method, Report Says » (TAVERNISE, 2015 ; l'italique a été ajouté par mes soins) où « top » fait explicitement référence à une RCT. Si le message est clair, il n'est pas pour autant tout à fait exact. Selon quelques spécialistes de la santé publique, trop d'attention a été accordée aux évaluations des traitements individuels au détriment de la recherche sur les systèmes de santé[32].

32. Voir, par exemple, RUTTER *et al.* (2017), Garchitorena *et al.* (chap. 5, ce volume).

Les plus grands *randomistas* ont également fait un bon travail de sensibilisation en apprenant aux autres à utiliser leur méthode de prédilection[33]. Les économistes du développement s'y sont rapidement mis, tout comme certaines ONG. Ils n'ont cessé de revoir à la hausse les critères qui définissent une bonne RCT, même si l'observation de HECKMAN et SMITH (1995), selon laquelle les RCT font l'objet d'un examen moins critique que les autres méthodes, semble toujours valable aujourd'hui.

Les fondateurs du J-PAL ont également fait part de leur désir de rendre le monde meilleur en élaborant des politiques fondées sur des preuves, ce qui a contribué à renforcer leur influence. Cette volonté avait été exprimée dès le départ par le J-PAL. Dans cette optique, le fait de réaliser beaucoup de RCT permettrait de caractériser ce qui fonctionne ou non, afin d'en tirer les enseignements nécessaires (BANERJEE, 2006). Une analogie est établie avec les RCT réalisées dans le cadre d'essais cliniques pour déterminer en moyenne quel médicament est le plus efficace (FAVEREAU, 2016).

Certains adeptes ont clairement été attirés par le zèle des grands *randomistas*. Partant de cela, « l'éthique expérimentale a été proposée comme le moyen de changer l'esprit du développement » (DONOVAN, 2018 : 27). Les *randomistas* peuvent être considérés comme un mouvement épistémique qui attire ses « vrais adeptes[34] » en défendant les RCT avec un « zèle quasi religieux » (Heckman, chap. 12, ce volume). Ce mouvement, qui défend bec et ongles les RCT, promet à ses adeptes une « révolution tranquille » (BANERJEE et DUFLO, 2011 : 265).

Les partisans (notamment les bailleurs de fonds) ont également été attirés par la simplicité des RCT, qui sont « plus transparentes et plus faciles à expliquer » (DUFLO, 2017 : 17). Les non-économistes comprennent plus facilement une RCT que les méthodes généralement privilégiées dans les études non expérimentales, qui devenaient de plus en plus sophistiquées et techniquement complexes au moment où le J-PAL a été créé.

L'influence des *randomistas* est-elle justifiée ?

Comme nous l'avons vu dans la partie précédente, les fondements statistiques ne nous indiquent pas que les RCT (lorsqu'elles s'y prêtent) sont nécessairement plus fiables, quel que soit le contexte, et trônent donc au sommet de la hiérarchie des méthodes. Il s'agit plus d'une question de foi que de science. Dans certains milieux, les méthodes qui ne sont pas basées sur la randomisation ont été clairement rejetées, ce qui témoigne d'une réaction excessive face aux difficultés rencontrées pour identifier les effets de causalité de cette façon.

33. La formidable « boîte à outils des RCT », conçue par DUFLO, GLENNERSTER et KREMER (2011), en est un exemple. Le blog *Development Impact* de la Banque mondiale fournit un soutien méthodologique très utile pour la réalisation des RCT : https://blogs.worldbank.org/impactevaluations.

34. Un critique de LEIGH (2018) décrit ce dernier comme un « vrai adepte » et raconte ensuite les différents choix personnels que Leigh opère dans sa vie sur la base des résultats des RCT (WYDICK, 2018).

L'analogie avec les essais cliniques n'est pas non plus convaincante. Il n'est pas évident qu'il soit possible d'utiliser des RCT de type « boîte noire » pour déterminer ce qui fonctionne ou non dans le domaine du développement, compte tenu de la diversité des interventions et des contextes. Bien souvent, les arguments avancés en faveur des RCT manquent d'une justification économique claire pour expliquer l'intervention, ou d'une structure cohérente pour comprendre pourquoi elle pourrait fonctionner ou non (HECKMAN et SMITH, 1995).

Alors que les *randomistas* spécialisés dans le développement invoquaient les essais cliniques comme modèle, les chercheurs en médecine adoptaient un point de vue plus nuancé[35]. D'un côté, certaines publications récentes laissent entendre que les préoccupations passées concernant le biais des études observationnelles causale en matière de santé et de médecine étaient excessives. D'un autre côté, il semble désormais admis que les avantages liés à la suppression des biais systématiques doivent être mis en balance avec les coûts et les risques des RCT cliniques.

Cependant, quelle que soit la validité de ces points, il faut reconnaître que le contexte médical est différent. Les économistes (et autres spécialistes en sciences sociales) traitent des personnes (en tant qu'individus et groupes) vivant dans des environnements sociaux et/ou économiques au sein desquels elles sont susceptibles de présenter une plus grande hétérogénéité, et dans lesquels le champ d'action est certainement plus vaste que dans les essais cliniques. Bien souvent, nous ne savons pas grand-chose du contexte spécifique *a priori*.

Certains problèmes d'inférence plus profonds se cachent derrière les affirmations des *randomistas* – des problèmes connus des experts des deux camps, mais mal compris de manière plus générale. Il existe très certainement une forme d'hétérogénéité non observée dans les impacts du traitement. Les sources sont nombreuses, et comprennent à la fois les circonstances (« *circumstances* ») de l'individu (comme l'expérience passée avec le type d'intervention) et les efforts déployés par les agents (reflétant leurs convictions sur l'impact)[36]. Cette hétérogénéité soulève la question suivante : « pour qui est l'impact ? » ANGRIST, IMBENS et RUBIN (1996) y ont répondu en montrant que l'estimateur par VI donne l'impact moyen pour un sous-ensemble de sujets traités, à savoir les « compliers », dont le statut de traitement a été modifié par la répartition aléatoire[37].

Lors de l'estimation de l'impact moyen sur les sujets traités, la validité de l'attribution aléatoire comme VI pour répondre au *take up* sélectif peut être mise en doute en présence de réponses comportementales présentant une telle hétérogénéité non observée des impacts du traitement (HECKMAN et VYTLACIL, 2005 ; HECKMAN *et al.*, 2006). Les différences d'impact doivent alors être reléguées

35. On trouve des exemples des points suivants dans CONCATO *et al.* (2000), SILVERMAN (2009), BOTHWELL *et al.* (2016), et FRIEDEN (2017).

36. Concernant cette source, voir CHASSANG *et al.* (2012), qui étudient les conséquences sur la validité externe des RCT.

37. Voir également la discussion dans PEARL (2009, chap. 8).

dans le terme d'erreur de la régression, en interaction avec le *take up* sélectif de l'attribution aléatoire. Les unités avec un rendement du traitement élevé seront plus susceptibles d'y adhérer. L'effet d'interaction, qui apparaît ensuite dans le terme d'erreur doit être corrélé à l'attribution aléatoire. La restriction d'exclusion ne s'applique pas. Bien sûr, cela n'a pas d'importance si l'on ne veut que l'ITT.

Il est rarement facile d'identifier les impacts des programmes sociaux, que l'attribution soit aléatoire ou non. Supposons que les caractéristiques latentes qui renforcent l'impact au niveau individuel aient également une importance pour les résultats contrefactuels d'une RCT avec une observance sélective. Le choix de la méthode d'estimation dépend alors essentiellement du paramètre d'impact à l'étude, du type de programme évalué et des réactions comportementales à ce programme (comme le montre Ravallion, 2014). Si les facteurs latents conduisant à un meilleur rendement du traitement sont associés à des résultats contrefactuels plus faibles, alors « la solution VI » pour le traitement endogène peut être pire que la maladie. En effet, l'estimateur des MCO peut même être non biaisé, malgré le *take up* sélectif. Il est essentiel que les praticiens réfléchissent bien aux possibles réponses comportementales à des impacts hétérogènes dans chaque application, comme dans toute étude non expérimentale.

Dans la pratique, la conception des RCT peut également nuire à l'identification. Il arrive que l'attribution aléatoire se fasse sur des grappes d'individus, par exemple des villages. Certaines grappes reçoivent le traitement et d'autres non. Les sujets faisant partie d'une grappe de traitement sont libres de suivre le traitement comme bon leur semble. Il s'agit d'une conception désormais classique dans le domaine du développement[38]. Mais elle se heurte à un problème chaque fois qu'il y a une interférence au sein des sous-groupes, de sorte que les non-participants des sous-groupes de traitement sélectionnés sont affectés par le programme. Par exemple, pour la RCT en grappes, Ravallion *et al.* (2015) ont montré un film didactique pour informer les gens de leurs droits en vertu de la loi indienne sur la garantie de l'emploi rural. Il était impossible de faire respecter une attribution des billets à certains individus au sein des villages, puisque le film devait être projeté dans des lieux publics, souvent ouverts, du village. L'accès au film a donc été attribué de manière aléatoire entre les villages, les habitants étant libres d'aller le voir ou non. Certains n'ont pas assisté à la projection, mais ils pouvaient (bien entendu) en discuter avec d'autres qui avaient vu le film, ce qui s'est avéré avoir un impact important sur les connaissances. La randomisation par grappes a dû être combinée à un modèle comportemental expliquant pourquoi certaines personnes ont regardé le film (Alik-Lagrange et Ravallion, 2019). Ce n'est qu'à partir de là que l'effet direct du traitement

38. Bien sûr, si l'on peut utiliser la double randomisation – aussi bien au sein des villages qu'entre eux –, on peut alors facilement traiter ce type d'interférence (Baird *et al.*, 2017). Les randomisations en grappes sont conçues pour les situations dans lesquelles la randomisation à l'intérieur d'une grappe n'est pas possible. Ce type de situation est courant dans le domaine du développement.

(regarder le film) a pu être isolé de l'effet indirect (vivre dans un village ayant accès au film). Dans cet exemple, les effets d'entraînement au sein des grappes enfreignent la restriction d'exclusion, et le fait d'utiliser l'attribution par grappes comme VI pour rendre compte du taux de *take up* individuel n'est donc pas très efficace.

Contrairement aux affirmations sur l'identification précise de l'impact causal moyen à l'aide d'une répartition aléatoire, des hypothèses et des modèles sont généralement nécessaires dans la pratique. Les hypothèses comportementales qui sous-tendent les études randomisées ne sont pas toujours explicites, ce qui ne facilite pas les choses (KEANE, 2010). En revanche, les approches structurelles poussent à davantage de clarté.

Certaines préoccupations ont suscité moins d'attention dans la littérature qu'elles ne le méritent. On peut citer l'exemple de l'effet Hawthorne, selon lequel le suivi modifie le comportement. Par exemple, si vous savez que vous faites partie du groupe de contrôle, vous pourriez avoir tendance à chercher un traitement de substitution. Ou bien certains sujets du groupe de traitement pourraient chercher à faire plaisir à l'expérimentateur[39]. Les RCT en économie sont rarement en double aveugle, contrairement aux essais cliniques, c'est pourquoi les biais associés au suivi sont plus fréquents, et ils méritent une plus grande attention dans le domaine du développement[40]. Un deuxième exemple est l'objet de la section suivante.

Prendre les objections éthiques au sérieux

Les préoccupations éthiques ne sont jamais très éloignées de l'élaboration des politiques. Ne pas prendre au sérieux la dimension éthique des évaluations présente deux dangers. Premièrement, il se peut que des évaluations jugées inacceptables sur le plan moral soient réalisées, le plus souvent dans des régions pauvres, où les populations sont vulnérables et où les institutions chargées de protéger leurs droits sont fragiles. Deuxièmement, cela risque de faire obstacle à des évaluations utiles sur le plan social, car trop risquées sur le plan politique, en grande partie par ignorance des avantages qu'elles présentent.

Pour certains détracteurs des RCT, « les randomisateurs sont prêts à sacrifier le bien-être des participants pour "tirer des enseignements" de l'étude » (ZILIAK

39. Au cours d'une RCT, la connaissance de l'intention de l'expérimentateur a été attribuée de façon aléatoire, mais aucun effet significatif n'a été constaté (MUMMOLO et PETERSON, 2019). D'autres tests de ce type sont nécessaires.

40. Cet aspect de la différence entre les RCT en économie et les RCT cliniques est abordé plus en détail dans FAVEREAU (2016). Pour en savoir plus sur l'effet Hawthorne dans le domaine de la santé, voir FRIEDMAN et GOKUL (2014).

et TEATHER-POSADAS, 2016)[41]. Ils ont par ailleurs souvent souligné que, dans une RCT, certaines personnes qui ont besoin du traitement ne le reçoivent pas, tandis que d'autres bénéficient d'un traitement dont elles n'ont pas besoin. Certains dénoncent également le fait que, dans les pays pauvres, les RCT ne font pas l'objet du même examen éthique attendu (sans être garanti) que dans les pays riches[42]. Dans certains essais cliniques randomisés menés dans des pays en développement et portant sur des traitements potentiellement dangereux, il a été rapporté que des participants n'étaient pas du tout conscients des risques sanitaires auxquels ils s'exposaient s'ils étaient traités[43]. BAELE (2013) soutient que les *randomistas* du développement ne se sont pas suffisamment souciés de la dimension éthique de leurs RCT. Face aux détracteurs des RCT, GLENNERSTER et POWERS (2016) avancent une argumentation éthique prudente.

La question de la validité éthique ne se pose pas de la même manière pour toutes les évaluations. Parfois, une évaluation d'impact s'appuie sur un programme existant, si bien que rien ne change dans le fonctionnement du programme. L'évaluation tient pour acquis les bénéfices du programme. Ainsi, si celui-ci est jugé éthique, on peut en déduire que l'évaluation l'est aussi. On peut qualifier ces évaluations « d'acceptables sur le plan éthique ».

D'autres évaluations d'impact modifient délibérément le mécanisme d'attribution (connu ou probable) du programme – qui bénéficie du programme et qui n'en bénéficie pas. De ce fait, le caractère éthique de l'intervention telle qu'elle est mise en œuvre à grande échelle n'implique pas nécessairement que l'évaluation soit éthique, elle aussi. On dit alors que ces évaluations sont « contestables sur le plan éthique », et les RCT en sont les principaux exemples dans la pratique. Les programmes à grande échelle ne s'appuient presque jamais sur une attribution aléatoire, de sorte que les RCT ont un autre mécanisme d'attribution, avec des bénéfices potentiellement très différents, étant donné l'hétérogénéité des impacts. Ainsi, la dimension éthique d'une RCT peut être contestée alors que le programme lui-même est satisfaisant à cet égard.

Il est sans doute assez extrême (position assez rare chez les économistes) de dire que les bonnes fins ne justifient jamais les mauvais moyens. Du point de vue éthique, il est défendable de juger les processus en partie en se basant sur leurs résultats ; en effet, l'idée selon laquelle les conséquences l'emportent souvent sur les processus – l'utilitarisme en étant le principal exemple – est une idée ancienne et répandue en philosophie morale. En soi, mener une RCT n'est pas contraire à l'éthique tant que celle-ci est motivée par les avantages qui peuvent découler des nouvelles connaissances. Toutefois, les avantages qui en

41. Voir également les commentaires de BARRETT et CARTER (2010), BAELE (2013), et MULLIGAN (2014).

42. Aux États-Unis, la dimension éthique de l'utilisation des RCT pour l'évaluation des politiques sociales fédérales n'a pas reçu la même attention que pour les essais cliniques. BLUSTEIN (2005) en expose les raisons.

43. Voir, par exemple, SATHYAMALA (2019) sur une RCT visant à étudier les risques sanitaires d'un contraceptif en Afrique.

découlent doivent être soigneusement soupesés par rapport aux risques liés au processus. Cela est particulièrement vrai dans les (nombreux) cas où il est possible de réaliser une étude non expérimentale acceptable sur le plan éthique.

La question de l'éthique a fait l'objet de nombreuses discussions dans la recherche médicale, où le principe d'équipoise exige qu'il n'y ait pas de cas antérieur déterminant laissant penser que le traitement aurait un impact[44]. Nous ne devons procéder à une randomisation ou continuer une RCT que si nous ignorons s'il est préférable d'être dans le groupe de traitement ou de contrôle[45]. Si les évaluateurs prennent au sérieux la validité éthique, certaines RCT de développement seront jugées inacceptables et donc rejetées, puisque nous sommes déjà plutôt sûrs des résultats et que les avantages tirés des connaissances ne seront probablement pas assez importants pour justifier une recherche contestable sur le plan éthique[46].

Le principe d'équipoise est rarement appliqué aux RCT menées dans le domaine du développement et des politiques sociales. D'ailleurs, c'est plutôt le contraire qui semble se produire. Dans un récent appel à propositions, un important bailleur de fonds philanthropique a explicitement exprimé sa préférence pour toute proposition de RCT « qui s'appuie sur des preuves préalables très prometteuses, suggérant qu'elle pourrait produire des impacts importants sur les résultats… » (Laura and John Arnold Foundation, 2018 : 2). D'un côté, on peut comprendre la préférence du bailleur de fonds, étant donné que les RCT coûtent cher et qu'il y a un désir d'avoir un impact avec des ressources limitées. Certains filtres *ex ante* de ce type sont sensés. On ne voudrait pas financer une RCT pour une intervention qui a peu de chances de se concrétiser sur le terrain. L'exemple ci-dessus met toutefois en évidence une certaine contradiction entre les objectifs des bailleurs de fonds et les préoccupations éthiques. La certitude *ex ante* d'un « impact considérable sur les résultats » nous fait craindre de priver de traitement ceux qui en ont besoin (et de le gaspiller en le donnant à ceux qui n'en ont pas besoin). Cela révèle également une inquiétude par rapport aux processus de financement qui déterminent ce qui est évalué. Nous reviendrons sur ce sujet dans la partie suivante.

Certains ont défendu le caractère éthique des RCT. Parmi les arguments avancés, on peut citer celui affirmant que les RCT sont justifiées chaque fois qu'un rationnement est nécessaire : lorsqu'il n'y a pas assez d'argent pour couvrir tout le monde, l'attribution aléatoire apparaît comme une solution équitable[47]. Cela a du sens lorsque l'on dispose de très peu d'informations. Parfois, dans le

44. Il y a une discussion intéressante chez Freedman (1987), qui a introduit le principe d'équilibre dans les essais cliniques. Dans le contexte des évaluations de l'impact sur le développement, voir Baele (2013) et McKenzie (2013).

45. Ici, le « nous » renvoie plutôt à un ensemble de personnes ayant une bonne connaissance de la littérature et une expérience pertinente. C'est ce que l'on appelle parfois « l'équipoise communautaire »

46. Voir les exemples évoqués dans Barrett et Carter (2010), Ziliak et Teather-Posadas (2016) et Narita (2018).

47. Voir, par exemple, le rapport de Goldberg (2014) sur les commentaires de Mulligan (2014). La même remarque est formulée par Fiennes (2018).

domaine du développement, nous en savons très peu *ex ante* sur la meilleure façon de répartir les participants afin de maximiser l'impact. Néanmoins, lorsque d'autres attributions sont possibles et que l'on dispose d'informations préalables sur les personnes susceptibles d'en bénéficier, il est sans doute plus juste de se servir de ces informations, et de ne pas procéder à une attribution aléatoire, du moins pas de manière systématique.

Il a également été avancé que la méthode de la randomisation conditionnelle (également appelée randomisation par « bloc » ou par « strate ») peut dissiper les préoccupations d'ordre éthique. Il s'agit de sélectionner d'abord les types de participants éligibles en fonction de la connaissance préalable des gains probables, et seulement ensuite d'attribuer aléatoirement l'intervention, étant donné que tous ne peuvent pas être couverts. Par exemple, si l'on évalue un programme de formation ou un programme qui exige des compétences pour obtenir un impact maximal, on peut raisonnablement supposer (en s'appuyant sur certaines preuves) que l'éducation et/ou l'expérience antérieure renforceront l'impact, et concevoir ensuite l'évaluation en conséquence. Cette méthode présente des avantages éthiques par rapport à la randomisation pure lorsqu'il existe des antécédents sur les impacts probables.

Mais il y a un piège. Ce qui est observable par l'évaluateur n'est généralement qu'un sous-ensemble des éléments visibles sur le terrain. Au niveau d'un village (par exemple), il y aura souvent plus d'informations que celles dont dispose l'évaluateur – des informations indiquant au niveau local que le programme est attribué à des personnes qui n'en ont pas besoin, et vice versa. Mais quelles informations devraient permettre de trancher ? Prétendre ne pas savoir serait une piètre excuse pour un évaluateur lorsque les autres parties prenantes savent très bien qui est dans le besoin et qui ne l'est pas.

Il a également été avancé que les modèles d'attribution aléatoire sont moins problématiques sur le plan éthique. L'idée est de n'interdire à personne l'accès à la prestation principale, mais plutôt de randomiser l'accès à une forme quelconque d'incitation ou d'information. Cela ne résout pas la question éthique, mais la déplace simplement de la prestation objet de l'étude vers un autre domaine. La validité éthique reste une préoccupation lorsque l'incitation est délibérément refusée à certaines personnes à qui elle profiterait et donnée à d'autres qui n'en ont pas l'utilité.

Prenons, par exemple, la RCT de Bertrand *et al.* (2007). Une des branches du traitement offrait une récompense financière importante aux participants qui obtenaient rapidement leur permis de conduire à Delhi, en Inde, ce qui les encourageait à verser des pots-de-vin aux fonctionnaires chargés de la délivrance des permis. Cette RCT ne permettait pas de verser directement des pots-de-vin ni de délivrer des permis à des personnes qui ne savaient manifestement pas conduire, mais ces résultats étaient prévisibles. Cette RCT devait permettre de vérifier que la corruption existe bel et bien en Inde et qu'elle a des effets réels. Cependant, il ne semble pas y avoir eu de doute sérieux préalable quant à la véracité de cette affirmation.

Les RCT peuvent être conçues pour aider à répondre aux préoccupations éthiques. Une option consiste à utiliser un « essai d'équivalence » dans le cadre duquel le groupe de contrôle reçoit ce qui est considéré comme le meilleur traitement suivant[48]. Contrairement aux contextes biomédicaux, il est possible que les avis soient partagés sur la *meilleure* option pour chaque application de développement spécifique. Néanmoins, il semble peu probable que les contrôles de type « ne rien faire » ou placebo, couramment utilisés, soient jugés acceptables lors d'un examen éthique minutieux dans la plupart des applications de développement. Il existe généralement une alternative. L'option « ne rien faire » n'est pas non plus susceptible de constituer un contrefactuel particulièrement pertinent pour la plupart des décideurs politiques.

Autre option : la randomisation adaptative. Elle est possible lorsque l'attribution est séquencée et que les réponses sont observées à chaque étape. Cette méthode permet de modifier l'attribution en cours de route, en fonction des preuves recueillies sur les impacts[49]. NARITA (2018) a proposé un plan adaptatif intéressant, semblable à celui du marché pour les expériences sociales, qui tient compte de la volonté de chaque participant de payer pour pouvoir bénéficier du traitement, compte tenu des connaissances préalables sur ses effets[50]. Contrairement à une RCT classique, on aboutit à une expérience d'optimum de Pareto[51], mais avec des propriétés statistiques similaires pour les estimations d'impact. Au moment où je rédige ce chapitre, cette idée ne semble pas avoir été mise en œuvre sur le terrain.

Aux États-Unis et ailleurs, les Institutional Review Boards (IRB) sont devenus courants lorsqu'il y a des propositions d'études sur des sujets humains. La plupart des institutions de recherche disposent d'un IRB spécifique. Ils sont en grande partie autorégulés. À quelques exceptions près, les processus des IRB appliqués dans le cadre de RCT de développement ne semblent pas avoir fait l'objet d'une évaluation systématique de leur efficacité. Une chose est sûre, les IRB doivent se concentrer davantage sur l'évaluation des bénéfices attendus d'une évaluation éthiquement contestable en tenant compte des connaissances existantes. Il existe des synthèses des connaissances actuelles qui peuvent être utiles et qui sont de plus en plus courantes[52].

48. Cette idée a été largement débattue dans le domaine biomédical, notamment dans le cadre des révisions apportées en 2000 à la déclaration d'Helsinki de 1964 par l'Association médicale mondiale. Pour plus de détails, voir LEVINE (2006).

49. La recherche biomédicale s'intéresse sérieusement à ces questions. Par exemple, la Food and Drug Administration (2010) a publié des lignes directrices des évaluations adaptatives. Voir aussi COX et REID (2000 : chap. 3).

50. Voir également CHASSANG *et al.* (2012) et la discussion dans ÖZLER (2018).

51. Un optimum de Pareto est une allocation de ressources sans alternative, c'est-à-dire que tous les agents économiques sont dans une situation telle qu'il est impossible d'améliorer le sort de l'un d'entre eux sans réduire la satisfaction d'un autre.

52. On les appelle parfois « revues systématiques » ; voir par exemple la base de données de 3ie et la Campbell Collaboration sur ces revues.

Si on les bouscule un peu, bon nombre de *randomistas* reconnaissent les préoccupations éthiques évoquées ci-dessus, même s'ils ne leur accordent que peu d'attention dans leurs articles. Ils partent du principe (le plus souvent implicite) que les avantages générés par leurs RCT compensent ces préoccupations. Il n'est pas toujours évident de savoir si cela est vrai ou non, et ce point mérite une plus grande attention.

Nous devrions également nous demander dans quelle mesure les efforts de recherche compensent les lacunes en matière de connaissances. Ces déséquilibres soulèvent d'autres préoccupations éthiques, compte tenu des défis pressants en matière de développement et des ressources limitées pour la recherche. La section suivante aborde ces questions.

Pertinence pour l'élaboration des politiques publiques

S'il est évident que des preuves solides ne suffisent absolument pas pour élaborer de bonnes politiques, les décideurs politiques se tournent de plus en plus vers ces éléments de preuve empirique dans l'espoir d'éclairer leurs choix et de remporter des débats politiques. La pertinence de la recherche évaluative en termes de politiques publiques constitue donc un enjeu important.

À ma connaissance, l'influence de toutes ces RCT sur la politique de développement n'a pas encore fait l'objet d'une évaluation complète et objective. On peut néanmoins citer des exemples de recherches pertinentes pour les politiques qui ont recours aux RCT. Prenons par exemple le cas de Banerjee *et al.* (2015a), qui ont mené des RCT dans six pays (Éthiopie, Ghana, Honduras, Inde, Pakistan et Pérou) afin d'évaluer l'approche mise en place depuis longtemps par le Building Resources Across Communities (BRAC) pour lutter contre la pauvreté en utilisant à la fois des transferts (actifs et espèces) ciblant les plus pauvres, avec une formation en matière d'alphabétisation et de renforcement des compétences[53]. Selon les chercheurs, l'adoption de l'approche du BRAC a permis de réaliser des gains économiques environ trois ans après le transfert initial d'actifs et un an après la fin des versements. En supposant que l'on extrapole ces gains dans un avenir lointain – ce qui est clairement une hypothèse forte –, leur valeur actuelle dépasse souvent le coût du programme semblable à celui du BRAC (Banerjee *et al.*, 2015a).

Sans prétendre à un examen exhaustif, cette réflexion met en évidence, en s'appuyant sur la littérature, certaines limites des RCT lorsqu'il s'agit de guider les politiques de développement.

53. L'ONG a débuté au Bangladesh (au départ sous le nom de Bangladesh Rural Advancement Committee), mais travaille maintenant dans de nombreux pays.

Paramètres pertinents

Même dans des conditions idéales, une RCT ne permet d'estimer qu'un sous-ensemble relativement limité de paramètres qui intéressent les décideurs politiques. En réalité, on s'attend à ce qu'il y ait à la fois des gagnants et des perdants, selon le contexte et les caractéristiques des unités participantes (et, comme on l'a vu, certaines de ces caractéristiques ne sont pas observées par l'analyste, bien qu'elles soient toujours des facteurs de motivation du comportement, notamment en ce qui concerne le choix d'adhérer au traitement ou non). Les impacts suivent donc une certaine distribution. Les décideurs politiques peuvent vouloir connaître la proportion des gagnants et des perdants au sein de la population, ou les types de personnes qui sont gagnants ou perdants. Pour identifier ces paramètres utiles à l'élaboration des politiques, il faut généralement disposer de plus de données et de méthodes économétriques structurelles. Un modèle structurel exhaustif n'est pas nécessairement essentiel pour traiter la question qui nous intéresse, mais (à l'inverse) une RCT fournira rarement les informations nécessaires.

Il existe des moyens fiables d'en apprendre davantage sur les impacts individuels plutôt que de se limiter à leur moyenne. Par exemple, l'estimateur local des variables instrumentales proposé par Heckman, Urzua, et Vytlacil (2006) vise à identifier les effets marginaux du traitement (*Marginal Treatment Effect* – MTE) pour toutes les valeurs de la probabilité empirique d'être traité. Contrairement à une RCT classique, les « essais sélectifs » permettent d'identifier les MTE en basant la probabilité d'attribution du traitement (plutôt que du contrôle) sur la volonté de payer exprimée par les agents (Chassang *et al.*, 2012). On peut ensuite regrouper les résultats pour obtenir l'impact moyen, comme avec une RCT. Mais on en apprend beaucoup plus que l'impact moyen.

Parfois, il est aussi possible de poser en toute fiabilité des questions contrefactuelles dans les enquêtes. C'est ce qu'ont fait Murgai *et al.* (2015), en demandant aux participants à un programme d'emploi ce qu'ils pensaient pouvoir toucher autrement (en comparant avec les résultats observés sur le marché du travail local). Cela permet d'en apprendre davantage sur la répartition des impacts, même si (bien sûr) il y a des erreurs de mesure dans les réponses aux enquêtes, et qu'il faudra donc très certainement faire une moyenne.

Pour juger de la performance, les décideurs politiques cherchent souvent à savoir qui bénéficie du programme, ce qui est (en partie) déterminé par le mécanisme d'attribution découlant de sa conception. S'il est déterminé par la demande, quelles sont les caractéristiques de ceux qui choisissent d'y adhérer ? S'il est rationné, à qui profite-t-il ? Ces questions se posent au premier stade des méthodes non expérimentales importantes qui utilisent l'appariement (*matching*) et qui commencent par un modèle statistique permettant de déterminer qui bénéficie ou non du programme[54]. Bien entendu, s'il s'agit d'une RCT, l'attribution

54. Il s'agit de l'appariement des coefficients de propension. Les valeurs prédites de ce modèle correspondent aux « scores de propension » utilisés pour sélectionner des groupes [suite p. suiv.]

n'est en principe pas prévisible, et si le programme est suivi de tous, on ne peut rien apprendre sur les types de personnes susceptibles d'y participer lorsque le programme sera mis en œuvre à plus grande échelle.

Dans le cas d'une observance imparfaite, nous pouvons en tirer des leçons grâce à la première étape de l'estimateur VI susmentionné. En effet, comme le soutiennent HECKMAN et PINTO (2019), une fois que nous admettons que le *take up* de l'attribution aléatoire est le résultat d'un choix rationnel, nous pouvons l'utiliser pour étudier à la fois les déterminants de la participation et identifier un plus large éventail d'autres paramètres. Par exemple, en faisant varier les incitations dans une RCT classique et en invoquant l'axiome faible des préférences révélées, les résultats de HECKMAN et PINTO (2019) peuvent être appliqués au problème du faible *take up* des politiques sociales, qui est assez fréquent chez les personnes pauvres et/ou socialement exclues. Au lieu de considérer l'observance sélective des sujets humains comme une nuisance statistique, nous pouvons en tirer des enseignements.

Une RCT peut également être utilisée pour évaluer l'impact probable *ex ante*, et ensuite faire une évaluation séparée du programme réel à l'échelle, en utilisant un estimateur basé sur l'observation. Cela semble prometteur, mais il faut comprendre que, compte tenu du *take up* sélectif et des impacts hétérogènes, on évalue essentiellement deux programmes différents, dont seulement un est effectivement mis en œuvre par le gouvernement. Il est facile de deviner lequel intéressera le plus les décideurs politiques. La deuxième évaluation sera-t-elle effectuée ? Il est possible que non si l'on adopte le point de vue de « l'étalon-or ».

La difficulté de cette évaluation de l'efficacité d'une politique réside dans le fait que la RCT est une construction plutôt artificielle, à la différence de presque toutes les politiques envisageables dans le monde réel.

Validité externe

Les décideurs politiques souhaitent évidemment tirer des leçons de ces expérimentations pour savoir si une intervention donnée peut fonctionner dans un autre contexte. Il s'agit de la question de la validité externe. Elle peut être contestée pour un certain nombre de raisons, notamment les effets de supervision (*monitoring*), les effets d'équilibre général, les problèmes d'échantillonnage et le soin spécifique apporté à administrer le traitement dans le cadre de la RCT (DUFLO, GLENNERSTER et KREMER, 2011).

Ces questions sont souvent ignorées dans la littérature relative aux RCT de développement, ou bien elles ne sont traitées que superficiellement. PETERS, LANGBEIN, et ROBERTS (2018) constatent que la majorité des 54 RCT dans le domaine du développement, publiées dans huit revues économiques (2009-2014),

de traitement et de comparaison équilibrés dans des études non expérimentales (ROSENBAUM et RUBIN, 1983).

n'abordent pas les raisons de l'invalidité externe et ne fournissent pas les informations nécessaires pour y remédier. Si plusieurs RCT portant sur une intervention donnée donnaient des résultats convergents, nous pourrions alors nous fier davantage à leur validité externe. Mais ce n'est pas le cas. VIVALT (2020) a documenté les écarts constatés dans les estimations d'impact d'un programme donné selon les contextes (et même les types d'évaluateurs). Ses conclusions mettent en garde contre les généralisations. Comme le fait également remarquer Vivalt, le manque de documentation sur les facteurs contextuels n'aide pas. Dans les exemples qu'ils donnent (pour les programmes de microcrédit), PRITCHETT et SANDEFUR (2015) montrent qu'une RCT (présumée) valide sur le plan interne et réalisée dans un contexte donné est inférieure à une étude non expérimentale permettant de prédire l'impact dans un autre contexte. Cette variabilité des estimations n'est pas entièrement due à l'hétérogénéité des impacts réels ; une estimation de sept RCT sur le microcrédit a montré que 60 % de la variabilité est due à la variation d'échantillonnage (MEAGER, 2019). En pratique, les décideurs politiques ont beaucoup de mal à distinguer la variation d'échantillonnage de la variabilité de l'impact réel.

Mener des RCT en collaboration avec des ONG, comme nous l'avons évoqué plus haut, présente des avantages qui ont également soulevé des questions quant à la validité externe. La RCT menée par DUFLO *et al.* (2015a) sur la scolarisation au Kenya en est un exemple. Des écoles choisies de manière aléatoire ont reçu les ressources nécessaires pour engager un enseignant supplémentaire avec un contrat à durée limitée. Les enfants ayant des enseignants contractuels ont obtenu de bien meilleurs résultats aux tests que ceux ayant des enseignants fonctionnaires. Cette expérience a été mise en œuvre par une ONG locale. Cependant, BOLD *et al.* (2018) ont tenté de reproduire cette expérimentation à grande échelle, en réalisant une RCT avec suivi, mais cette fois avec une branche mise en place par le gouvernement (ainsi qu'une autre par l'ONG). Il en est ressorti que l'amélioration des résultats des tests était en fait liée à la mise en œuvre par une ONG et non au type d'enseignant. L'effet « enseignant » constaté par DUFLO, DUPAS et KREMER (2015) avait disparu.

Une estimation en forme réduite (qu'elle provienne d'une RCT ou non) est une « boîte noire » qui n'apporte pas beaucoup d'éléments pour répondre aux objectifs de l'élaboration des politiques. Tirer des leçons des RCT pose des problèmes spécifiques. Réfléchir à la manière dont nous pourrions tirer des enseignements d'une RCT sur le passage d'une intervention à une plus grande échelle constitue un objectif important. Une RCT mélange de manière aléatoire des personnes à faible impact (pour lesquelles les bénéfices attendus du programme sont faibles) avec des personnes à fort impact, sur la base de caractéristiques latentes. Le programme mis en œuvre à grande échelle comptera vraisemblablement davantage de personnes à fort impact, qui seront attirées par le programme[55]. Compte

55. C'est un exemple de ce que HECKMAN et SMITH (1995) ont surnommé « biais de randomisation ». Voir également la discussion de Heckman (chap. 12, ce volume), qui revient sur cette question à la suite de l'essor des RCT dans le domaine du développement.

tenu de cette sélection intentionnelle basée sur les impacts attendus (hétérogènes), le programme national est fondamentalement différent de celui de la RCT, qui ne contient sans doute que peu d'informations utiles pour évaluer le programme à grande échelle.

Cela reflète une observation plus générale de MOFFITT (2006), selon laquelle de nombreuses choses peuvent changer – les *inputs* et même le programme lui-même – lors de la mise à l'échelle d'un projet pilote. Une ONG qui souhaite attirer des bailleurs de fonds aura tout intérêt à montrer l'impact d'un essai qui n'est pas caractéristique de ses opérations habituelles. Les jeunes chercheurs qui effectuent un essai sur le terrain sont susceptibles de déployer davantage d'efforts que les responsables gouvernementaux qui mettent en œuvre la version à grande échelle. La validité externe impose des contraintes sur la conception et l'exécution des essais pilotes qui ne sont pas suffisamment prises en compte dans la pratique.

Pour en savoir plus sur la validité externe, une approche consiste à répéter l'évaluation dans différents contextes. Par exemple, GALASSO et RAVALLION (2005) ont étudié, à l'aide d'une méthode non expérimentale, les performances du programme bengali « *Food-for-Education* » dans 100 villages et ont corrélé les résultats avec les caractéristiques de ces villages. Les différences de performance étaient en partie imputables aux caractéristiques observables dans les villages, telles que l'inégalité de répartition des terres à l'intérieur des villages (les villages les plus inégaux parvenant moins bien à atteindre les pauvres). Dans les évaluations précédentes, le fait de ne pas tenir compte de ces différences a été considéré comme une véritable lacune[56]. Regarder à l'intérieur de la boîte noire d'une évaluation d'impact peut apporter un éclairage utile sur sa validité externe et ses implications politiques. Pour ce faire, il faut généralement disposer d'informations extérieures à la conception initiale de l'évaluation. On peut citer comme exemple la RCT *Proempleo* de GALASSO *et al.* (2004). D'un côté, des personnes participant à un programme d'emploi ont reçu de manière aléatoire des coupons pour des subventions salariales, de l'autre il y avait le groupe de contrôle randomisé. En théorie, la subvention salariale permet de réduire le coût de la main-d'œuvre pour l'entreprise et rend donc l'embauche plus attrayante. Conformément aux prévisions de cette théorie, la RCT a eu un impact significatif sur l'emploi. Toutefois, une vérification ultérieure des registres administratifs a révélé un recours très faible des entreprises aux subventions salariales. *Proempleo* n'a donc pas fonctionné comme le prévoyait la théorie. Des entretiens qualitatifs de suivi avec les entreprises et les travailleurs ont indiqué que les coupons avaient une valeur de certification pour les travailleurs – une sorte de « lettre de recommandation » réservée à peu de personnes (le fait qu'elle ait été attribuée de manière aléatoire n'ayant pas été signalé au niveau local dans cette RCT). Cela n'a pas été révélé grâce à la RCT, mais a nécessité des données d'observation supplémentaires. Cela n'avait pas été prévu par les chercheurs *ex ante*.

56. Voir par exemple les commentaires de MOFFITT (2004) sur les essais de réformes de l'aide sociale aux États-Unis.

D'ailleurs, s'ils s'en étaient tenus strictement au plan de pré-analyse, ils seraient passés à côté d'un élément crucial, pertinent pour les politiques, qui explique l'impact du programme. Les données supplémentaires ont également révélé l'importance de fournir des informations sur la façon d'obtenir un emploi, ce qui a eu des répercussions sur la transposition du programme à plus grande échelle. Un passage à grande échelle de la subvention salariale en se basant sur la seule RCT aurait été une erreur.

Un pan de la littérature a utilisé la randomisation (soit de l'intervention, soit d'un déterminant clé de son placement) pour mettre en lumière des paramètres structurels plus profonds. Cela a été le cas dans certaines des premières applications en matière d'évaluation de la politique sociale aux États-Unis (Heckman, chap. 12, ce volume). Dans un exemple d'applications récentes en matière de développement, Todd et Wolpin (2006) utilisent la RCT dont j'ai parlé précédemment pour le projet *Progresa* au Mexique afin de modéliser les réponses comportementales dynamiques à l'incitation à la scolarisation fournie par ce programme. Ces recherches peuvent nous aider à comprendre les impacts d'un programme et faciliter les simulations de politiques alternatives. Todd et Wolpin montrent que l'extension de la subvention de *Progresa* vers des niveaux de scolarisation plus élevés permettrait d'améliorer l'impact global. Dans le même ordre d'idées, il est possible d'utiliser une RCT pour tester un ou plusieurs liens clés de la « théorie du changement » qui sous-tend la logique d'un programme, même si l'outil n'est pas applicable au programme lui-même. Cela fait écho aux arguments de Heckman (1992) et de Heckman et Pinto (2019) sur la possibilité de mener des expériences plus ambitieuses fondées sur la théorie.

Lacunes en matière de connaissances

Pour faciliter l'élaboration des politiques de lutte contre la pauvreté, les chercheurs devraient idéalement combler les écarts entre nos connaissances sur l'efficacité des politiques et les connaissances nécessaires aux décideurs politiques. De toute évidence, les choses ne se passent pas comme nous pourrions l'espérer. Par exemple, Kapur (2018) rapporte un entretien avec Arvind Subramanian, ancien conseiller économique en chef du gouvernement indien (*Government of India* – GOI) : « Lorsqu'on lui a demandé, parmi toutes ces RCT coûteuses, combien avaient fait avancer la politique en Inde, Arvind Subramanian […] a eu du mal à en trouver ne serait-ce qu'une seule qui lui ait été utile pour résoudre les dizaines de questions politiques urgentes auxquelles il était confronté[57]. »

Pourquoi y a-t-il de telles lacunes dans les connaissances ? Il existe des facteurs aléatoires, mais aussi des « défaillances du marché du savoir » plus systématiques (Ravallion, 2009b). L'une des raisons de cette situation est la présence d'externalités dans les évaluations. Certains éléments portent à croire que la réalisation

57. Voir aussi les commentaires de Basu (2014), un autre ancien conseiller économique en chef du gouvernement indien.

d'une évaluation d'impact d'un projet de développement en cours peut contribuer à améliorer certains aspects de sa mise en œuvre, notamment en accélérant le déblocage des fonds (Legovini *et al.*, 2015). Cependant, les enseignements tirés d'une évaluation sont également utiles pour les projets futurs, qui (avec un peu d'espoir) profiteront des leçons tirées des évaluations précédentes. On ne peut pas attendre des chefs de projet actuels qu'ils tiennent bien compte de ces avantages externes lorsqu'ils décident du budget à consacrer à l'évaluation de leur propre projet. Les externalités varient considérablement selon les types d'évaluation, et sont nettement plus importantes lorsque les évaluations sont plus innovantes, voire les premières du genre. Les externalités des évaluations jouent également un rôle dans le « biais de myopie » dans le domaine du développement, si bien que les évaluations à long terme sont rares (Ravallion, 2009b ; Bouguen *et al.*, 2019).

Les lacunes du marché des connaissances découlent également des biais de publication liés à la fois aux processus de sélection des éditeurs de revues et au comportement des auteurs, notamment en matière de documentation de leurs résultats. Les résultats négatifs ou nuls ont moins de chances d'être publiées ou même de faire l'objet d'une quelconque mention[58]. Les réplications ultérieures des expériences en économie trouvent souvent des effets moins importants[59]. Dans certains cas, les résultats antérieurs ont été reproduits correctement, mais ils se sont révélés très sensibles à des aspects discutables de l'analyse des données, qui n'étaient pas évidents dans l'article original[60].

La dynamique des processus de publication explique également la persistance des lacunes en matière de connaissances. La littérature n'est pas exempte d'erreurs et il faut parfois du temps pour les corriger. Compte tenu de son originalité, le premier article sur un sujet donné a de grandes chances d'être publié et de bénéficier d'une grande visibilité. Les articles suivants auront tendance à être relégués dans des revues de moindre importance, à être moins souvent cités, ou même à ne pas être publiés du tout. L'auteur de l'article original devient alors le gardien des connaissances sur le sujet. On peut parfois réussir à le contourner, mais il exerce toujours une influence considérable. Cependant, il se peut que le premier article soit erroné. De plus, les mesures visant à encourager les efforts de réplication semblent faibles en économie[61]. Pourtant, en

58. Sur 221 études en sciences sociales, il a été constaté que « les résultats positifs ont 40 % plus de chances d'être publiés que les résultats négatifs et 60 % plus de chances de faire l'objet d'un compte rendu » (Franco *et al.*, 2014 : 1502). La distribution des valeurs *p* rapportées dans les articles publiés dans l'*American Economic Review* (AER), le *Quarterly Journal of Economics* (QJE) et le *Journal of Political Economy* suggère que les chercheurs ont tendance à faire des choix de spécification qui gonflent la significativité de leurs résultats pour franchir la barre des « 5 % » (Brodeur *et al.*, 2016). Christensen et Miguel (2018) étudient les caractéristiques de ces biais dans les publications de recherche économique et discutent de la manière dont ceux-ci pourraient être réduits.

59. Camerer *et al.* (2016) ont reproduit 18 expériences de laboratoire publiées dans l'AER et le QJE. En moyenne, l'effet reproduit était inférieur d'un tiers à l'effet original.

60. Voir, par exemple, Bédécarrats *et al.* (2019a), qui met en doute la validité interne et externe de la RCT originale de Crépon *et al.* (2015).

61. Voir la discussion de Rodrik (2009). Depuis lors, le 3ie a soutenu les efforts de réplication des évaluations d'impact dans le domaine du développement par le biais de sa Replication Window et son *Journal of Development Effectiveness*.

sciences, les échecs de réplication sont fréquents (IOANNIDIS, 2005a). Le premier constat de la distribution des impacts peut donc avoir un effet de distorsion durable sur les connaissances reconnues.

L'invalidité externe soulève également des inquiétudes quant au processus d'accumulation des connaissances. Même si le premier article s'avère proche de la vérité dans le contexte spécifique, il peut avoir une validité limitée dans d'autres circonstances. Lorsque le sujet concerne l'impact d'une politique, ou une problématique très pertinente pour l'impact en question, les connaissances sur les politiques auront tendance à être faussées en conséquence.

Il s'agit là de préoccupations d'ordre général, qui ne se limitent pas aux RCT. Cependant, la hiérarchie des méthodes basée sur « l'étalon-or » pourrait bien aggraver les choses, comme nous allons le voir maintenant.

Adapter les efforts de recherche aux défis des politiques publiques

Les lacunes en matière de connaissances sont également dues à un décalage des efforts d'évaluation. Les évaluateurs spécialisés dans le développement ignorent trop souvent les possibilités de *fongibilité*. Les bénéficiaires (gouvernementaux ou non) peuvent réaffecter leurs propres efforts pour répondre à de nouveaux fonds, comme l'aide au développement. Les bailleurs de fonds financent souvent autre chose de manière implicite. Même si cette conséquence est moins connue, il se peut que les bailleurs de fonds et les niveaux supérieurs de gouvernement apprécient mal leur propre impact : ils évaluent le projet que le bénéficiaire de l'aide a proposé de financer plutôt que le projet qui a été réellement financé, compte tenu des possibilités de fongibilité. Les efforts d'évaluation ne sont alors pas en phase avec les efforts de développement.

S'il ne faut pas incriminer les RCT, les préférences méthodologiques marquées des évaluateurs risquent toutefois de renforcer ce déséquilibre. Les *randomistas* du développement ont eu à la fois des effets de production et de substitution des connaissances. La multiplication des RCT depuis 2000 (fig. 1) suggère au moins un effet de production positif. Toutefois, comme nous l'avons déjà évoqué, les validités interne et externe de ces RCT dans le champ du développement ne sont pas tout à fait évidentes. Nous ignorons le contrefactuel – ce que nous aurions appris si ces ressources (financières et humaines) avaient été déployées ailleurs.

L'effet de substitution est lié aux méthodes utilisées. Prenons l'exemple de la Banque mondiale. Alors que la première RCT enregistrée dans la base de données de 3ie a été réalisée par la Banque mondiale, jusqu'au début des années 2000, cet outil n'était considéré que comme l'une des nombreuses options fiables en matière d'évaluation d'impact. Depuis, la Banque mondiale s'est nettement orientée vers les RCT, ce qui a été salué par certains analystes. Par exemple, dans un éditorial de *The Lancet*, on peut lire (dans la plus grande ignorance de l'histoire) que « la Banque mondiale se tourne enfin vers la science » (Lancet,

2004 : 731)[62]. L'Independant Evaluation Group (IEG) de la Banque mondiale rapporte que plus de 80 % des évaluations d'impact réalisées entre 2007 et 2010 ont utilisé la randomisation, contre 57 % en 2005-2006 et seulement 19 % les années précédentes (World Bank, 2012).

Même si nous supposons que toutes ces RCT ont eu un effet positif sur les connaissances, l'effet de substitution pourrait bien fonctionner dans l'autre sens. Ce dernier comporte trois aspects. Premièrement, le fait de mettre l'accent sur l'identification des effets causaux par le biais des RCT a détourné l'attention des autres méthodes d'investigation empirique, notamment de la recherche descriptive, qui est clairement sous-estimée aujourd'hui dans la recherche sur le développement. Certains des enseignements politiques tirés d'articles de recherche basés sur les RCT auraient pu provenir de bonnes « descriptions robustes » (« *thick descriptions* », utilisant des méthodes qualitatives et/ou quantitatives) des processus du monde réel liant les interventions aux résultats.

Deuxièmement, on craint que l'accent mis sur les programmes individualisés attribués n'ait détourné l'attention de la recherche systémique, qui recourt généralement à des modèles structurels. En économie, plus largement, les travaux structurels dans l'enseignement et la recherche ont perdu de leur importance, comme l'ont noté Keane (2010) et d'autres. Ce problème est également une source de préoccupation importante pour la recherche sur la santé publique (Rutter *et al.*, 2017).

Troisièmement, l'évaluation d'impact du portefeuille de politiques de développement pose un problème, car la randomisation n'est possible que pour un sous-ensemble non aléatoire de politiques et de contextes. Il devient donc impossible de faire des inférences sur un large éventail de politiques si l'on se fie uniquement aux RCT. La randomisation est généralement mieux adaptée aux programmes dont les participants et les non-participants sont clairement identifiés, dont la durée est relativement courte, qui n'exigent pas l'imposition de redevances/taxes et dont les coûts ou les avantages peuvent difficilement bénéficier au groupe de non-participants. Les RCT sont donc plus adaptées aux biens privés, qui sont faciles à attribuer parmi les ménages individuels, qu'aux biens publics dont les avantages sont partagés entre de nombreuses personnes (Hammer, 2017). Il existe des exceptions (comme pour certains biens publics locaux). Toutefois, il est généralement beaucoup plus difficile de randomiser le lieu des projets d'infrastructure de moyenne ou grande ampleur et il semble impossible de randomiser les réformes sectorielles et macro-économiques. Cet outil n'est donc utile que pour un nombre limité de politiques essentielles à la stratégie de développement d'un pays.

Pour évaluer l'impact de la fourniture de biens privés, il faut une justification économique de cette « politique ». Les marchés ne pourraient-ils pas fournir le bien privé de manière efficace, sans devoir procéder à une évaluation d'impact ?

62. Sur l'influence des RCT à la Banque mondiale, voir Webber et Prouse (2018).

Il peut y avoir de bonnes raisons pour justifier la réalisation d'une évaluation d'un bien privé dans un contexte spécifique, mais le plus souvent, il semble que les *randomistas* ne font que poursuivre les opportunités de randomisation. Certes, les objectifs de redistribution sont parfois évoqués, mais de manière assez superficielle. Les impacts distributifs (sur la pauvreté, par exemple) sont rarement abordés avec rigueur, ni même identifiés comme des résultats explicites. Autrement dit, l'économie publique est souvent absente.

Prenons l'exemple de la déforestation dans les pays en développement pour illustrer en quoi le recours intensif aux RCT fausse les connaissances nécessaires à l'élaboration des politiques. Il est fréquent que les ménages qui possèdent des bois et coupent leurs arbres ne tiennent pas compte du coût externe de leur contribution au réchauffement climatique. Il existe une solution connue depuis longtemps : la taxe pigouvienne. Mais elle serait difficile à mettre en œuvre par le biais d'une RCT, puisque la taxation est principalement du ressort des gouvernements, qui seraient (naturellement) réticents. Au lieu de cela, on peut randomiser les indemnités versées à ceux qui choisissent de ne pas abattre leurs arbres, comme c'est le cas avec la RCT menée en Ouganda par JAYACHANDRAN *et al.* (2017). Cette politique peut être mise en œuvre par une ONG locale, en contournant le gouvernement. Il y a ici une justification d'économie publique, mais l'utilisation d'une RCT limite les options de politique évaluées. En particulier, la politique fiscale aura probablement des impacts différents (ne serait-ce que parce que la politique de versement donne une valeur supplémentaire au stock d'arbres, générant un effet revenu, distinct de l'effet prix).

Bien entendu, aucun outil ne peut à lui seul convenir à toutes les situations. Reste à savoir si nous avons trouvé aujourd'hui un équilibre raisonnable entre les efforts de recherche et les défis politiques. La hiérarchie (discutable) des méthodes préconisées par les *randomistas* rend cet équilibre plus difficile à atteindre. En effet, aussi pour les biens privés, l'idée même de l'attribution aléatoire est contraire aux objectifs de nombreux programmes de développement, qui visent généralement à toucher certains types de personnes ou de lieux. En offrant des transferts monétaires aux pauvres – type d'intervention très prisé des RCT dans le domaine du développement –, les gouvernements pourront, si tout va bien, obtenir de meilleurs résultats qu'avec une attribution aléatoire.

Le rapport de l'IEG déjà cité décrit le déséquilibre dans la répartition des évaluations d'impact de la Banque mondiale entre les différents secteurs de ses opérations, ainsi que le manque apparent de cohérence entre le portefeuille d'évaluation et les priorités sectorielles et de développement de la Banque mondiale (World Bank, 2012). Même si je n'ai pas de preuves, je soupçonne également un déséquilibre entre les efforts d'évaluation selon la durée estimée des bénéfices du projet. Les évaluations à long terme des projets de développement de la Banque mondiale sont rares, malgré les arguments avancés concernant les effets à long terme. Mon expérience personnelle m'a montré combien il est difficile d'organiser et de mettre en œuvre des évaluations à long terme à

la Banque mondiale[63]. Il y a tout lieu de penser que le fait de privilégier les RCT exacerbe le biais de myopie des connaissances en matière de développement.

Cela n'a pas seulement lieu à la Banque mondiale. Le biais sectoriel observé dans les RCT au sens large est évident d'après les résultats de Cameron *et al.* (2016), qui présentent un tableau croisé de plus de 2 200 évaluations d'impact publiées (dans la base de données de 3ie susmentionnée) par méthode et par secteur[64]. Au total, les deux tiers environ de ces évaluations ont recours aux RCT, mais ces dernières ont tendance à se limiter à certains secteurs, notamment l'éducation (58 % ont utilisé une RCT), la santé, la nutrition et la population (83 % ; 93 % pour la santé seule), les technologies de l'information et de la communication (67 %), et l'eau et l'assainissement (72 %). Les études non expérimentales sont plus fréquentes – moins d'un tiers des évaluations utilisent une RCT – dans les domaines de l'agriculture et du développement rural, de la politique économique, de l'énergie, de l'environnement et de la gestion des catastrophes, du développement du secteur privé, des transports et du développement urbain. La réalisation d'évaluations d'impact est également inégale sur le plan géographique (même si l'on tient compte de la population). L'Inde en a enregistré le plus grand nombre absolu, mais le Kenya le plus grand nombre par habitant[65]. La répartition géographique des RCT est influencée par les relations qu'entretiennent les chercheurs avec les ONG locales.

Ce biais se manifeste à la fois du côté de l'offre et de la demande. Du côté de l'offre des évaluations, on constate aujourd'hui que de nombreux économistes et autres spécialistes des sciences sociales et politiques, séduits par la promesse d'identifier clairement un effet de causalité, cherchent quelque chose à randomiser. Si la randomisation n'est pas possible, ils se tournent vers une autre question.

Du côté de la demande, les gouvernements (et les agences de développement) sont libres, en règle générale, dans leur choix des sujets à évaluer. Ils ne savent néanmoins pas toujours de quel type de preuves ils ont besoin (Duflo, 2017), ce qui est préoccupant. La politique joue également un rôle. Ils peuvent être tentés de choisir des programmes pour lesquels une évaluation négative ne risque pas, ou peu, de leur nuire politiquement, ou de choisir ceux qui sont réellement importants, mais pour lesquels il y a de bonnes raisons de croire à un résultat politiquement satisfaisant (ce qui soulève à nouveau des préoccupations d'ordre éthique). Les autres programmes importants ne seront pas évalués. Les risques sont évidents.

Pour répondre à ces préoccupations, il faut des agendas d'évaluation plus stratégiques, qui ne soient pas motivés par les préférences méthodologiques des chercheurs. C'est ce que nous commençons à voir pour les RCT. Si l'on peut s'en réjouir, les stratégies sont néanmoins toujours menées par des chercheurs

63. Ceci est largement basé sur l'étude rapportée par Chen *et al.* (2009).

64. En plus des RCT, voici les autres méthodes identifiées : les différences de différences, les variables instrumentales, les régressions par discontinuité et l'appariement. Plusieurs méthodes sont admises dans les comptages.

65. Pour plus de détails, voir Cameron *et al.* (2016) et Sabet et Brown (2018).

universitaires, en fonction de leurs intérêts et consacrées à un seul outil. Si nous sommes vraiment soucieux d'obtenir des estimations d'impact non biaisées du portefeuille des politiques de développement, il vaudrait certainement mieux choisir avec soin (ou peut-être même au hasard !) ce qui est évalué, et ensuite trouver la meilleure méthode pour les programmes sélectionnés, en considérant les RCT comme une option parmi d'autres. Il s'agit là d'une obligation si nous avons vraiment à cœur d'obtenir une évaluation impartiale de l'impact global sur le développement. Si la recherche peut servir cet objectif, cela ne se fera sûrement pas automatiquement.

Conclusion

Nous assistons à une évolution bienvenue vers une culture de l'expérimentation pour lutter contre la pauvreté et relever d'autres défis en matière de développement. Les RCT figurent au menu des outils existants à cet effet. Toutefois, elles ne méritent pas le statut privilégié que leur ont accordé leurs défenseurs, et qui a tant influencé les chercheurs, les agences de développement, les bailleurs de fonds et l'ensemble de la communauté du développement. Pour pouvoir établir un classement fiable de deux modèles d'évaluation, il ne suffit pas de savoir que l'un d'entre eux a recours à la randomisation.

La popularité des RCT repose sur une prétendue hiérarchie des méthodes, avec au sommet les RCT faisant figure « d'étalon-or ». Lorsqu'on l'examine de près, cette hiérarchie ne tient pas la route. Malgré les nombreuses revendications contraires, une RCT ne met *pas* sur un pied d'égalité les résultats contrefactuels entre les unités traitées et les unités de contrôle. L'absence de biais systématique ne signifie pas que l'erreur expérimentale d'une RCT ponctuelle est inférieure à l'erreur d'une autre méthode non aléatoire. Il est impossible de le savoir. Parmi toutes les méthodes envisageables pour un projet (avec un budget donné pour l'évaluation), l'option de la RCT n'est pas nécessairement celle qui va se rapprocher au plus près de la vérité. En effet, si la taille de l'échantillon d'une étude non expérimentale est supérieure à celle d'une RCT dans un même contexte, alors les essais par étude non expérimentale se rapprochent plus souvent de la vérité, même en présence de biais.

Il existe encore un champ des possibles assez vaste pour réaliser des études non expérimentales, ainsi que d'autres études non randomisées utiles (telles que des assignations expérimentales déterministes), fondées sur la théorie. Si la modélisation fait l'objet d'incertitudes, elles ne sont généralement pas aussi grandes que ne le laissent entendre les *randomistas*. De plus, lorsque nous examinons les RCT en pratique, nous constatons qu'elles sont confrontées à des problèmes d'erreurs de mesure, d'observance sélective et de contamination. Il apparaît alors évident que cet outil ne peut pas répondre aux questions que nous posons

sur la pauvreté et les politiques de lutte contre celle-ci sans faire le même type d'hypothèses que celles que l'on trouve dans les études non expérimentales, hypothèses que les *randomistas* entendent éviter.

Les RCT sont également contestables d'un point de vue éthique, notamment concernant certaines dimensions qui touchent moins les études non expérimentales. Il est impossible d'interpréter la dimension éthique des RCT sans évaluer les bénéfices attendus des nouvelles connaissances, au regard de ce qui est déjà connu. Les commissions d'examen doivent accorder plus d'attention à la situation *ex ante*, où une intervention est délibérément refusée à ceux qui en ont besoin, et délibérément accordée à ceux qui n'en ont pas besoin, dans un but d'apprentissage. Dans certains contextes spécifiques, les arguments peuvent être valables, compte tenu des limites des connaissances existantes, mais ils doivent être présentés de manière crédible et non pas simplement tenus pour acquis.

Les affirmations controversées concernant la supériorité des RCT comme étant « l'étalon-or » ont faussé l'utilisation des évaluations d'impact destinées à éclairer l'élaboration des politiques de développement. Ce biais provient du fait que la randomisation n'est possible que pour un sous-ensemble non aléatoire de politiques. Lorsqu'un programme s'étend à l'ensemble d'une communauté ou d'une économie, ou lorsqu'il y a des effets de contamination ou d'entraînement entre les personnes traitées et celles qui ne le sont pas, une RCT n'apportera pas grand-chose, et pourrait bien s'avérer trompeuse. Cet outil ne convient qu'à un éventail assez restreint de politiques de développement, et même dans ce cas, il ne permet pas de répondre à bon nombre des questions que se posent les décideurs politiques. Défendre les RCT comme étant la meilleure, voire la seule, méthode scientifique d'évaluation d'impact risque de fausser nos connaissances en matière de lutte contre la pauvreté. Ce risque était l'une des principales préoccupations de RAVALLION (2009a), et l'expérience n'a fait que la renforcer.

Malgré les grands progrès réalisés au cours des dix dernières années, il y a encore des raisons de douter que la recherche évaluative sur le développement corresponde bien aux défis politiques actuels. Ce chapitre a démontré que, pour un meilleur alignement, il faut :

– abandonner les revendications selon lesquelles il existe une hiérarchie inconditionnelle des méthodes avec au sommet les RCT, et montrer clairement que les preuves « scientifiques » et « rigoureuses » ne se limitent pas aux RCT ;

– exiger une déclaration *ex ante* claire et bien documentée des bénéfices attendus d'une RCT, à mettre en balance avec des compromis éthiques douteux ;

– expliciter les hypothèses comportementales qui sous-tendent les évaluations randomisées, à l'instar des règles des approches structurelles ;

– aller au-delà des impacts causaux moyens, pour inclure d'autres paramètres d'intérêt politique et mieux comprendre les mécanismes qui lient les interventions aux résultats ;

– considérer les RCT comme un outil parmi d'autres pour remédier aux lacunes de connaissances relatives au portefeuille des politiques de développement.

Remerciements

François Roubaud a encouragé l'auteur à rédiger ce chapitre. L'auteur remercie Jason Abaluck, Sarah Baird, Radu Ban, Mary Ann Bronson, Caitlin Brown, Sylvain Chabé-Ferret, Kevin Donovan, Ryan Edwards, Markus Goldstein, Miguel Hernan, Emmanuel Jimenez, Max Kasy, Madhulika Khanna, Nishtha Kochhar, Agnès Labrousse, Andrew Leigh, David McKenzie, Rachael Meager, Berk Özler, Dina Pomeranz, Lant Pritchett, Milan Thomas, Vinod Thomas, Eva Vivalt, Dominique van de Walle, Andrew Zeitlin, ainsi que les personnes ayant participé à un atelier d'auteurs à Paris en mars 2019. Le personnel de l'International Initiative for Impact Evaluation qui a bien voulu fournir une mise à jour de sa base de données sur les évaluations d'impact publiées et répondre aux questions de l'auteur.

Chapitre 2

Randomiser le développement

Méthode ou pure folie ?

Lant Pritchett

Introduction

Bill Gates faisait l'éloge récemment de la possession de volailles pour combattre la pauvreté en Afrique. Dans une lettre ouverte, le professeur Chris Blattman, de l'université de Chicago, a fait remarquer que des transferts monétaires pourraient s'avérer plus rentables que des poulets : « Il serait assez simple de réaliser une étude sur quelques milliers de personnes dans six pays, en utilisant huit ou douze variantes, pour comprendre quelle combinaison est la plus efficace, où et avec quelle population. *Cette réponse me semble être le meilleur investissement que nous puissions faire pour lutter contre la pauvreté dans le monde.* Les experts de l'organisation Innovations for Poverty Action (IPA), qui ont mené l'essai sur les animaux d'élevage pour la revue *Science*, sont d'accord avec moi. En fait, nous avons tenté, ensemble, de lancer une étude comparative de cette nature[1]. »

Je pense qu'à ce stade, il est important pour la communauté du développement, que nous, membres de celle-ci, nous arrêtions un instant pour nous demander comment nous en sommes arrivés à cette double folie. Celle, d'abord, qui peut inciter Bill Gates – génie, figure humanitaire et grande personnalité publique intellectuelle – à être un tant soit peu sérieux lorsqu'il parle de ses fameux « poulets ». Celle de la méthode, ensuite, qui peut conduire Chris Blattman – génie lui aussi, représentant d'une université de premier plan mondial, et non moins grande figure intellectuelle publique – à répondre non pas « Des poulets ? Vous y songez vraiment ? », mais plutôt que le « meilleur investissement » pour « lutter contre la pauvreté dans le monde » consiste à mettre en œuvre la *bonne*

1. https://www.cgdev.org/blog/getting-kinky-chickens (c'est l'auteur qui souligne).

méthode pour étudier les programmes concurrents de la fourniture de poulets, c'est-à-dire par exemple celui des transferts monétaires, et en concevoir les éléments adaptés[2].

J'espère qu'il est évident qu'il s'agit bien là de folie. Les 20 premiers pays en développement les plus peuplés dans le monde sont (dans l'ordre) : la Chine, l'Inde, l'Indonésie, le Brésil, le Pakistan, le Nigeria, le Bangladesh, la Russie, le Mexique, les Philippines, l'Éthiopie, le Vietnam, l'Égypte, l'Iran, la Turquie, la République démocratique du Congo, la Thaïlande, l'Afrique du Sud, la Tanzanie et la Colombie. Ils représentent à eux tous 4,6 milliards d'habitants. Imaginez maintenant que vous rassembliez quelques dizaines de leaders de l'un de ces pays (leur *leadership* pouvant être politique, social, économique, intellectuel, populaire, issu d'un mouvement de masse, de la société civile, ou de plusieurs natures à la fois), et que vous leur disiez : « Nous, experts de la communauté du développement, pensons que la "lutte contre la pauvreté mondiale" est au cœur du programme de développement, et que le "meilleur investissement" possible pour promouvoir le développement/combattre la pauvreté de votre pays [remplacer par : Indonésie, Brésil, Nigeria, République démocratique du Congo, Tanzanie, Afrique du Sud, Égypte, Inde] est de mener une série d'études utilisant la bonne méthode, afin de déterminer si les programmes contre la pauvreté doivent promouvoir la possession de volailles ou distribuer de l'argent, et, dans ce contexte, définir comment concevoir au mieux ces programmes de fourniture de poulets ou de transferts monétaires. »

J'imagine deux réponses que pourraient faire les chefs de file des pays concernés. La première pourrait être : « comment avez-vous pu en arriver à des idées aussi triviales et dévalorisantes sur les objectifs, ambitions et défis de notre pays ? Comment, en tant qu'[Indonésiens/Indiens/Nigérians/Égyptiens/Tanzaniens, etc.], ne pas prendre comme un véritable outrage le fait de suggérer que des "poulets" ou "des études sur les poulets" puissent faire partie des priorités absolues de notre pays ? » La seconde réponse envisageable serait : « nous pouvons facilement répertorier, pour votre gouverne, de nombreux enjeux de développement pressants, urgents, voire critiques pour le bien-être actuel et futur des citoyens de notre pays. Ces questions sont importantes, que votre méthode préférée pour fabriquer des articles de recherche puisse ou non y apporter des réponses[3]. »

2. Avec des dizaines d'études sur les transferts monétaires conditionnels, la microfinance, et l'apparition du qualificatif « *Worm Wars* » pour décrire le débat de masse sur la rentabilité des campagnes de vermifugation (sans oublier un ensemble de RCT sur des interventions anti-pauvreté de petite envergure et à objectifs fétichistes), cette folie s'est insinuée bien plus largement encore.

3. Voici quatre anecdotes (parmi les nombreuses possibles) pour étayer cette allégation. La première a été vécue par l'un de mes collègues dans le bureau du Premier ministre d'un grand pays. À la demande d'éminents *randomistas* qui avaient réalisé un travail considérable dans ce pays, il avait réussi à organiser une rencontre de deux heures entre ces universitaires et le Premier ministre. À la fin de l'entretien, le Premier ministre a pris à part mon ami et lui a dit : « Ne me faites plus jamais perdre mon temps ainsi. » Deuxième anecdote : mon collègue Arvind Subramanian a été pendant trois ans l'un des principaux conseillers stratégiques de l'Inde, pays qui a été le centre d'attention des *randomistas*. Dans une intervention devant mes étudiants en 2018, il a expliqué qu'au cours des trois années pendant lesquelles il avait participé à différents niveaux (des niveaux [suite p. suiv.]

Lorsque j'utilise l'expression « études poulets contre transferts monétaires », je ne vise pas spécifiquement le professeur Blattman, mais plutôt l'ensemble du mouvement des *randomistas* qui sévissent dans le domaine du développement. Au lieu de trouver difficile de penser à « autre chose » (LUCAS, 1988) qu'aux enjeux généraux liés au développement national, les économistes du développement sont aujourd'hui tellement attachés à la méthode qu'ils pensent « à tout sauf » au développement national. Il existe actuellement des milliers d'évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) qui ont fait l'objet d'une publication, dont des dizaines d'études sur les transferts monétaires conditionnels et la microfinance, et des centaines d'autres sur des actions de petite envergure mises en œuvre dans les domaines de l'eau, de l'assainissement, de l'enseignement, de la santé, de la formation en gestion, etc. Ce que je veux souligner, c'est que toute cette folie au sujet de la méthode employée dans les travaux académiques sur le développement n'est qu'un symptôme, et non la maladie elle-même. Le grand débat concerne l'importance relative du « développement national » par rapport à un « développement fétichiste » ou étroit (*kinky development*), et consiste à savoir s'il est possible d'accélérer le « développement national ». L'analyse randomisée utilisée en tant que méthode ne peut prétendre à une quelconque importance que (a) si l'on interprète le développement dans le sens restreint de la réalisation d'objectifs très étroits et de faible niveau (le développement « fétichiste »), ou (b) si l'on est d'avis que le « développement national » échappe totalement à l'influence des idées ou des preuves.

Le « développement national » est la quadruple transformation d'un groupe intrinsèquement social (un pays, une région ou une société) vers des capacités de niveaux supérieurs dans quatre dimensions : une transformation économique pour augmenter la productivité ; une transformation politique pour rendre les gouvernements plus sensibles et réactifs aux vastes attentes de la population ; une transformation administrative pour accroître les capacités fonctionnelles d'exécution des organisations (y compris celles qui relèvent de l'État) ; et une transformation sociale pour un traitement plus égalitaire des citoyens du pays (habituellement accompagnée d'un sentiment d'identité commune et, dans une

intermédiaires aux niveaux les plus éminents) à des débats sur les divers défis économiques rencontrés par l'Inde, il n'avait *jamais* entendu dire que les résultats de RCT avaient joué un quelconque rôle. Troisième exemple : dans le cadre de mes travaux d'expert du développement, je me suis rendu dans la quasi-totalité de ces 20 premiers pays les plus peuplés (sauf deux), et j'ai vécu plusieurs années dans deux d'entre eux (l'Indonésie et l'Inde), et je n'ai jamais entendu les poulets ou les études rigoureuses être cités comme des priorités hors du cercle étroit des agences et des projets de développement. Quatrièmement, lorsque l'essai sur les animaux d'élevage publié dans *Science* a été mis en avant dans les médias, une journaliste d'un magazine américain m'a appelé pour savoir ce que je pensais de cette étude importante. Je lui ai répondu que je ne l'avais pas lue, car en tant que chercheur et spécialiste du développement, elle n'avait pas pour moi un caractère particulièrement intéressant ou important. Elle m'a demandé comment je pouvais dire cela, au vu de ses augustes auteurs et de la prééminence de la revue. Je lui ai répondu que si elle pouvait trouver la moindre mention de cette étude dans les médias locaux de l'*un* des sept pays considérés, je changerais d'avis, lirais l'étude et lui transmettrais mes commentaires. Comme, bien sûr, la journaliste ne m'a jamais rappelé, j'ai chargé un assistant de recherche de compulser les médias de chacun des pays pris en compte dans l'étude (en mettant à contribution des locuteurs natifs), et nous n'en avons pas trouvé une seule mention à l'échelle locale.

certaine mesure, d'un but commun). Pour des pays comme Haïti, l'Inde, la Bolivie ou l'Indonésie, le développement national consiste à atteindre les hauts niveaux de capacités fonctionnelles économiques, politiques, administratives et sociales du Danemark, du Japon ou encore de l'Australie. Le développement national n'est pas une fin en soi, mais un moyen d'atteindre un degré supérieur de bien-être humain.

Le « développement fétichiste » (PRITCHETT, 2014a ; KENNY et PRITCHETT, 2013) repose sur l'idée que le développement consiste principalement, pour ne pas dire exclusivement, à parvenir à des seuils minimaux sur des indicateurs spécifiques : « vaincre l'extrême pauvreté », « assurer l'enseignement primaire pour tous » ou « avoir accès à l'eau potable » sont des objectifs « fétichistes » dans le sens où ils tracent une ligne ou un seuil complètement arbitraire dans une dimension du bien-être humain, en prétendant ensuite qu'une déviation marginale de la distribution du bien-être qui amènerait la population justement à ce niveau minimal est l'objectif même du développement. L'élément distinctif de ce développement « fétichiste » est que les gains en matière de bien-être humain qui dépasseraient ce seuil minimal ne représentent plus rien.

Dans le développement de ce chapitre, la première partie démontrera empiriquement deux choses. Premièrement, cette partie démontre que le revenu médian/la consommation médiane, l'un des quatre éléments du développement national, est à la fois (a) empiriquement *nécessaire* et *suffisant(e)* pour réduire le taux de pauvreté de consommation (en pourcentage de la population), et (b) explique *l'essentiel* de la variation des taux de pauvreté entre les pays. En outre, les effets des programmes anti-pauvreté – *a fortiori* la conception de ces programmes et, « doublement *a fortiori* », les études sur cette conception – sont tout simplement infimes, comparés aux impacts de la croissance inclusive.

Deuxièmement, pour des indicateurs composites du bien-être humain, tel que l'indice de progrès social (*Social Progress Index* – SPI) : (a) que des niveaux élevés de développement national sont empiriquement nécessaires et suffisants pour atteindre des niveaux élevés de bien-être humain, et (b) que ce lien est empiriquement très étroit en ce qui concerne le SPI (et d'autres indicateurs du bien-être humain). De plus, la totalité de la douzaine de mesures spécifiques du bien-être humain (à une exception près) qui entrent dans l'indice de progrès social (par exemple l'accès à l'eau, la sécurité des personnes, la santé, l'éducation, etc.) est également fortement corrélée avec le développement national.

La deuxième partie présentera un cadre de type arbre de décision pour évaluer l'allégation selon laquelle une activité intellectuelle spécifique (comme une RCT) menée sur des programmes ciblés (du genre transferts monétaire contre fourniture de poulets) pourrait constituer le « meilleur investissement » pour « lutter contre la pauvreté » (ou, plus généralement, pour toute mesure en faveur du bien-être humain). Je montre que *toutes* les étapes du raisonnement qui sont nécessaires pour arriver à une telle conclusion sont erronées.

Développement national et bien-être humain

Je propose ci-après une définition très générale du « développement national » ainsi que des indicateurs empiriques de celui-ci, et je montrerai ensuite sa relation empirique avec des mesures du bien-être humain, qu'il s'agisse de mesures étroites « fétichistes », comme un seuil de pauvreté extrême, ou bien de mesures plus larges.

Le développement national : une transformation des pays dans quatre dimensions

Le terme même de « développement » implique un changement dans le temps par lequel quelque chose évolue vers une version plus mature, plus avancée et supérieure de sa nature fondamentale. Un être humain se développe de l'état d'embryon jusqu'à celui d'adulte, et la grenouille, avant de devenir telle, passe d'abord par les stades de zygote, puis de têtard. Mais les roches, elles, ne se « développent » pas pour devenir des grenouilles, pas plus qu'elles ne se « développent » pour devenir du sable du fait de l'érosion. Le premier cas est impossible, et le second ne correspond pas à une évolution « directionnelle ». Alors qu'est-ce qui se « développe » lors d'un « développement » ? Dans le cas du développement « national », le sujet qui se « développe » est typiquement un pays. Mais il s'agit toujours, intrinsèquement, d'un groupe social (et socialement construit)[4]. Un pays possède quatre dimensions importantes (au moins) à travers lesquelles il se « développe », chacune étant intrinsèquement et ontologiquement sociale et ne pouvant pas être réellement individualisée.

Développement économique

On entend généralement par cette expression la capacité productive d'un endroit. Elle comprend certains éléments caractéristiques des individus, mais possède également une dimension générale de type « productivité multifactorielle » qui est spécifique à l'endroit et non individualisée. L'un des indicateurs possibles du développement économique d'un pays est la productivité du travail, que l'on mesure par le PIB par travailleur, même s'il existe plusieurs autres méthodes (voir par exemple les méthodes de mesure de la complexité économique selon Hidalgo et Hausmann, 2009), le PIB pouvant en outre être ajusté de diverses façons (comptabilité verte, par exemple). Ces indicateurs ne sont *jamais* destinés à fournir des mesures directes du bien-être humain, mais constituent des mesures du produit et de la productivité économiques d'un pays.

4. Alors que les termes « nation » ou « État-nation » sont souvent utilisés négligemment comme synonymes de « pays », ils comportent un important bagage idéologique sur ce qu'est une « nation » et sur son rapport avec des États souverains comme des « pays ». Mais nous pouvons parler du « développement » de régions (par exemple l'Italie du Sud par rapport à l'Italie du Nord), ou de provinces/États au sein d'un pays (comme le Tamil Nadu par rapport à l'Uttar Pradesh).

Développement administratif

Par développement administratif, on considère habituellement qu'il s'agit d'un ensemble d'éléments représentant la capacité d'organisations (principalement étatiques) à réaliser des objectifs d'ordre public[5]. Les pays disposent d'un éventail d'organisations destinées à mener à bien des missions : les armées, les banques centrales, les services postaux, les forces de police, les tribunaux, les bureaux du cadastre, etc. Même si la capacité de ces organisations varie bien entendu au sein des pays (KAUFMANN *et al.*, 2002), la capacité administrative globale d'un État constitue un autre élément du développement national. L'indice des états fragiles (*Fragile States Index* – FSI), qui est un exemple d'indicateur pour cette dimension, note les pays de 0 (le meilleur score) à 10 (le moins bon score, attribué aux pays les plus fragiles) sur la base de leur « capacité générale à fournir des services publics », le Danemark obtenant une note de 0,9, l'Indonésie de 5,6 et Haïti de 9,4.

Développement politique

Cette expression est évidemment fortement connotée et, comme tout ce qui se dit concernant la politique, elle porte elle-même une valeur politique. Sur le plan descriptif toutefois, lorsque certaines personnes ont pu décrire le « développement » des États, elles avaient généralement à l'esprit l'idée que les détenteurs du pouvoir politique qui exercent l'autorité souveraine dans un pays : (a) répondent aux besoins, souhaits, attentes, volontés des citoyens de ce pays, sachant que des processus politiques permettent que ceux-ci soient exprimés par les citoyens et collectés par des moyens justes et légitimes, et (b) respectent au moins certains droits dits « négatifs » préservant la liberté et la sécurité des individus (et peut-être aussi certains droits « positifs ») ; à cela s'ajoute (c) un certain degré d'« État de droit ». Le FSI comprend par exemple deux indicateurs distincts, un pour la « légitimité de l'État » (et non la « démocratie ») et un pour « les droits de l'Homme et l'État de droit » (10 étant la note la plus mauvaise et 0 la meilleure). Concernant la légitimité de l'État, le score de Haïti est de 8,7, celui de l'Indonésie de 4,8 et celui du Danemark de 0,9, alors qu'en matière de droits de l'Homme et d'État de droit, il est de 7,4 pour Haïti, 7,3 pour l'Indonésie et 1,2 pour le Danemark. L'indicateur Polity2 du projet Polity IV va de 10 à moins 10, 10 représentant la démocratie totale et - 10 l'autocratie totale. Ainsi, le score du Danemark est de 10 depuis 1915 (avec l'interrègne de la Seconde Guerre mondiale) ; en Indonésie, il était de - 5 en 1998 (dernière année de présidence de Soeharto), grimpant à 6 en 1999, puis à 9 en 2017 ; en Haïti, il était de 0 de 2010 à 2015, et de 5 en 2016 et 2017.

5. Dans Andrews *et al.* (2017), nous établissons une distinction entre la *capacité* des organisations, qui est caractéristique d'une organisation, et l'*aptitude* en tant que trait distinctif des individus, et nous montrons que la capacité d'une organisation n'est pas la somme des aptitudes des individus. Ceci vise à souligner qu'il existe deux concepts distincts, mais nous reconnaissons que les termes pourraient aussi être échangés (en utilisant par exemple l'aptitude en tant que caractéristique d'une organisation) pour atteindre le même but, dans la mesure où la distinction entre ces deux concepts est observée de manière cohérente.

Développement social

La façon dont les citoyens/membres d'une même société se traitent mutuellement est encore plus connotée, et donc plus politique que le développement politique lui-même. C'est une notion variable qui appartient intrinsèquement au développement. Bien que la notion d'« égalité sociale » – dans le sens où les individus sont traités de manière égale par leurs congénères, quelle que soit leur identité sociale (famille, classe héréditaire, clan, tribu, appartenance ethnique, race, sexe, religion) – ait été altérée à bien des égards (émanant pour beaucoup de projections condamnables des constructions sociales issues des colonialistes et du colonialisme), elle fait par nature partie du développement. L'un des aspects du développement social a été la création/l'adoption d'une identité commune. Il s'agit évidemment de valeurs qui se sont construites historiquement en Occident, et qui ne sont pas universellement valables. Mais je dirais qu'elles ont souvent été intégrées, en bien ou en mal, dans les concepts de « modernisation » et de « développement ». Il est aujourd'hui particulièrement évident que le développement doit comporter une approche sexospécifique, et que les sociétés dans lesquelles la notion d'égalité des sexes n'existe pas sont considérées comme moins « développées socialement », au moins dans une dimension importante, que celles dans lesquelles elle est prise en compte.

Les « unités » au niveau desquelles se déroule le développement national, à savoir un marché, une organisation, un État, une société représentent des processus auxquels les individus prennent part et dans lesquels ils sont intégrés, sans toutefois y être individualisés sur le plan ontologique.

Les niveaux de revenu médian/consommation médiane expliquent pleinement la pauvreté

Le développement national – et, dans le cas présent, seulement un indicateur du développement national, à savoir les niveaux de consommation médiane – suffit pour éliminer (globalement) la pauvreté dite « extrême » (dont le seuil se situe aujourd'hui, avec l'inflation, à 1,90 P $ par jour[6]). Les données compilées par la Banque mondiale, limitées aux données d'enquête auprès des ménages de toutes les paires « pays/année », permettent de disposer de plus de 800 observations (pays/année) sur les taux de pauvreté et sur les revenus ou la consommation médians. La fig. 1 montre qu'aucun pays avec un revenu annuel médian supérieur à 3 000 P $ (soit environ le niveau du Pérou ou de la Mongolie autour de 2010) n'a un taux d'extrême pauvreté dépassant les 10 %. À 5 000 P $ (soit environ le niveau du Costa Rica), globalement aucun pays n'a un taux d'extrême pauvreté, supérieur à 2 %. En outre, aucun pays ayant un revenu annuel médian supérieur à 1 000 P $ (environ le niveau du Bangladesh en 2010) n'a un taux d'extrême pauvreté, supérieur à un tiers de sa population. L'espace vide situé dans la zone « nord-est » de la fig. 1 est important, car il représente des combinaisons revenu médian/taux de pauvreté que l'on ne rencontre jamais. Il existe donc un niveau

6. « P $ » signifie « dollar ajusté pour la parité de pouvoir d'achat [PPA] ».

de revenu médian/consommation médiane empiriquement suffisant pour réduire la pauvreté au-dessous d'un pourcentage donné de la population.

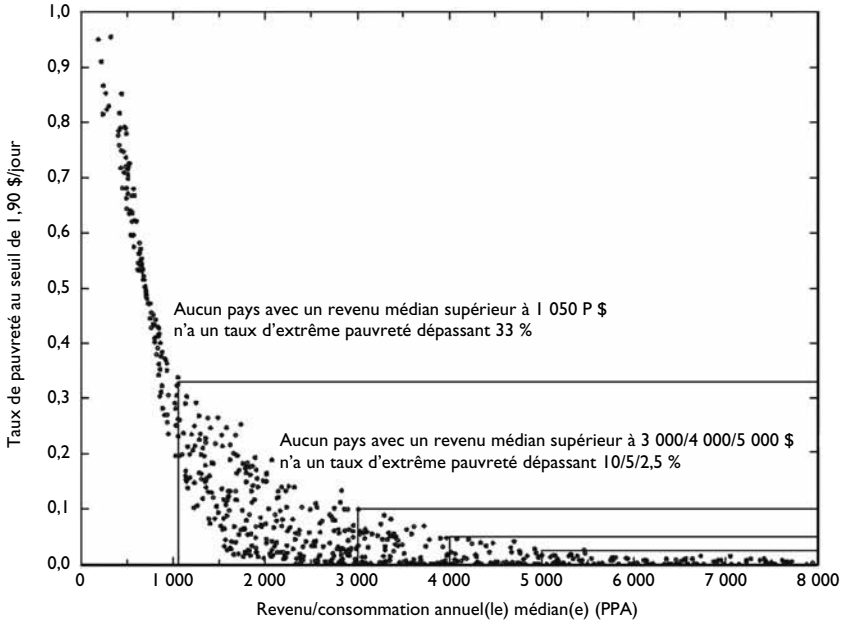


Figure 1

Le revenu médian/la consommation médiane est suffisant(e) pour éliminer l'extrême pauvreté.

Source : calculs réalisés par Lant Pritchett avec des données issues de PovcalNet ; l'outil de calcul en ligne pour la mesure de la pauvreté a été conçu par le Development Research Group de la Banque mondiale (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

La fig. 2 montre les niveaux de revenu médian/consommation médiane qui sont empiriquement nécessaires pour atteindre différents niveaux de taux de pauvreté à un seuil de 5,5 $ par jour[7]. Par « empiriquement nécessaires », je n'affirme pas qu'il s'agit d'une nécessité logique (comme un théorème), mais simplement que cela ne se produit pas. L'espace vide dans la zone « sud-ouest » de la fig. 2 correspond aux combinaisons de faible revenu médian/faible taux de pauvreté que l'on ne rencontre jamais. Aucun pays n'a fait baisser le taux de pauvreté à un seuil de 5,5 $ par jour en dessous de 75 % des ménages sans disposer d'un revenu médian de plus de 1 045 P $. Ceci signifie que, pour 42 des 164 pays, leur dernier niveau de revenu observé est tel qu'*aucun pays* n'a *jamais* enregistré un taux de pauvreté (au seuil de 5,5 P $) de moins de 75 % avec ce niveau de revenu. Et, pour 107 des 164 pays, le niveau de revenu est tel que (quasiment) aucun pays n'a enregistré un taux de pauvreté inférieur à 10 % avec ce niveau

7. C'est le seuil de pauvreté le plus élevé pour lequel la Banque mondiale fournit des données, mais il s'agit d'un seuil « modéré », et non d'un seuil « élevé ». Je préconise, comme de nombreux autres experts, de définir les seuils de pauvreté de tranche supérieure à 10 P $ par jour ou plus, soit des valeurs encore très inférieures à celles qui sont effectivement utilisées dans les pays plus riches.

de revenu. Aucun pays (sauf un[8]) n'a réussi à faire passer le taux de pauvreté – dont le seuil est de 5,5 P $ par jour – en dessous des 10 % sans bénéficier d'un revenu médian/d'une consommation médiane de plus de 3 535 P $, soit environ le niveau des pays à « revenu intermédiaire de la tranche supérieure », comme le Pérou (3 486 P $ en 2015), le Kazakhstan (3 557 P $ en 2015) ou la Thaïlande (3 549 P $ en 2010).

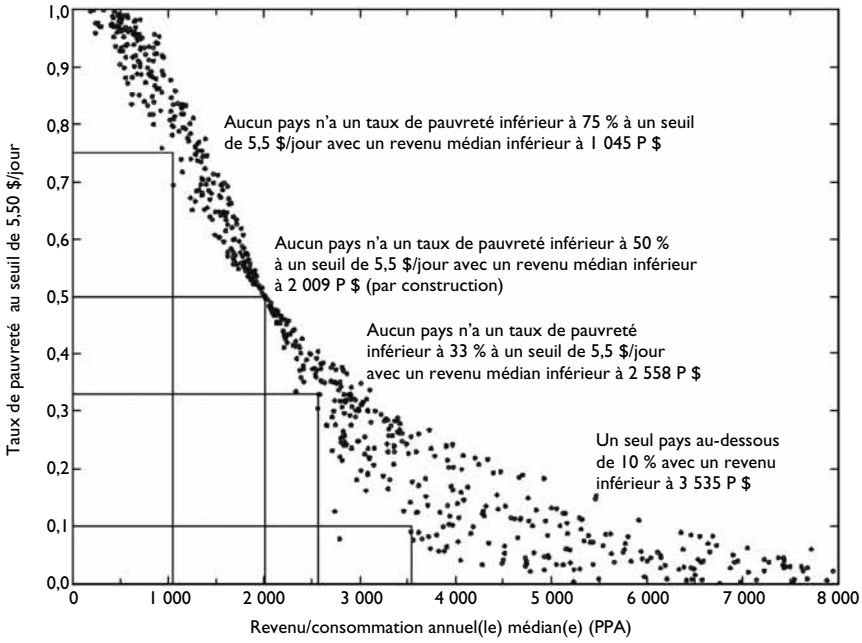


Figure 2

Des niveaux élevés de revenu médian/consommation médiane sont empiriquement nécessaires pour éliminer la pauvreté

(et ces niveaux sont d'autant plus élevés que le seuil de pauvreté est important).

Source : calculs réalisés par Lant Pritchett avec des données issues de PovcalNet ; l'outil de calcul en ligne pour la mesure de la pauvreté a été conçu par le Development Research Group de la Banque mondiale (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

J'ai utilisé jusqu'à présent 810 observations issues des données de la Banque mondiale, qu'elles concernent le revenu ou la consommation. Toutefois, pour étudier les relations avec les programmes ou projets, les dépenses de consommation constituent un meilleur indicateur, car elles mesurent de manière plus fiable les résultats après impôt et avec les transferts, et reflètent donc les dépenses de consommation en intégrant les éventuelles prestations versées dans le cadre des

8. Il s'agit de l'Azerbaïdjan en 2005, pour lequel les données font apparaître un revenu médian de 5 655 P $ et un taux de pauvreté de 5 % pour un seuil de 5,5 $ par jour en 1995, un revenu médian de 5 197 P $ et un taux de pauvreté quasi égal à zéro en 2015, mais un revenu médian de 2 785 P $ et un taux de pauvreté de 7,7 % en 2005, qui est une observation anormalement basse, même pour ce pays.

programmes. La fig. 3 montre la relation entre les taux de pauvreté à l'échelle nationale (pour les trois seuils de pauvreté de 1,9 P $, 3,2 P $ et 5,5 P $ observés par la Banque mondiale) et la médiane de la distribution de la consommation, en utilisant seulement les 389 observations pays/année qui intègrent des données de consommation. Étant donné que les taux de pauvreté doivent, par construction, être non linéaires à la médiane, j'ai ajusté une forme fonctionnelle entièrement flexible, incluant toutes les puissances de la médiane de - 2 à 5.

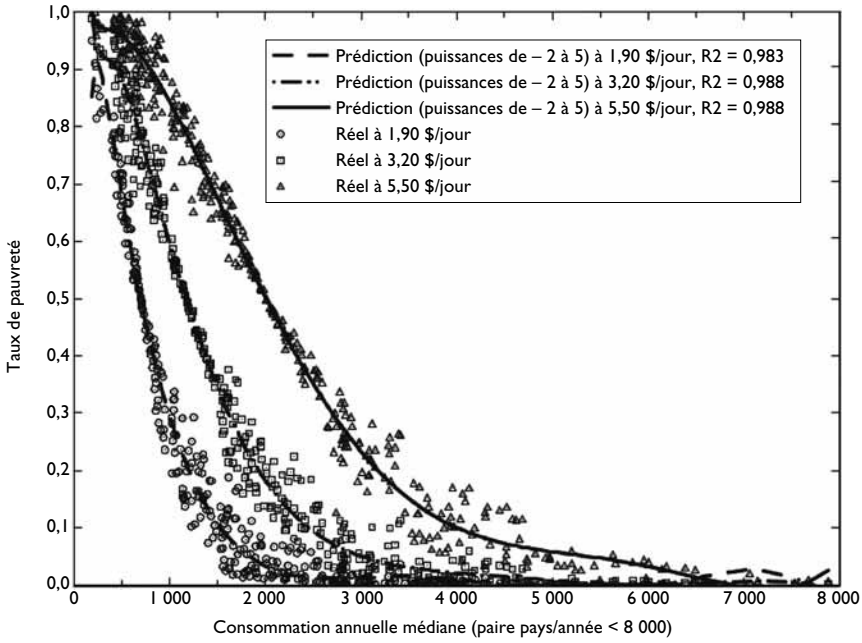


Figure 3

La consommation médiane d'un pays permet de prédire le niveau de pauvreté de manière exacte pour les seuils de pauvreté élevés, et quasi exacts pour les seuils de pauvreté inférieurs.

Source : calculs réalisés par Lant Pritchett avec des données issues de PovcalNet ; l'outil de calcul en ligne pour la mesure de la pauvreté a été conçu par le Development Research Group de la Banque mondiale (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

Pour l'ensemble des trois seuils, les données indiquent que la quasi-intégralité de la variation observée au niveau des taux de pauvreté entre les pays et dans le temps (avec un R2 allant de 0,983 à 0,988) est associée à une variation de la médiane de consommation (50^{e} centile). Une valeur R2 de 0,988 signifie que la corrélation entre les taux de pauvreté réels et le taux de pauvreté prédit à partir de la médiane atteint 0,994 (= $\sqrt{0,988}$).

Cela ne veut bien entendu pas dire que d'autres facteurs comme l'évolution des inégalités ou l'adoption de programmes « anti-pauvreté » ne peuvent pas avoir un effet, voire ne peuvent pas par principe jouer un rôle « substantiel », mais indique simplement que, sur le plan empirique, en comparaison avec les changements

massifs associés aux variations de la médiane (les taux de pauvreté variant de 100 % à presque 0 %), les différences induites à un certain niveau de consommation sont très modestes comparées aux bénéfices tirés de la croissance. Le tabl. 1 donne des calculs de différents contrefactuels de pauvreté. Pour un pays situé au milieu du quartile inférieur, le taux de pauvreté est de 72,2 %. Si le pays s'était déplacé « plein sud », c'est-à-dire s'il enregistrait un taux de pauvreté plus bas, pour la même consommation médiane, de la valeur d'un écart-type du résidu, le taux de pauvreté serait de 68,6 %. En revanche, si ce pays avait vu sa consommation médiane augmenter de 2 points par an sur les 20 années précédentes (soit environ la valeur d'un écart-type des taux de croissance transnationaux), son taux de pauvreté aurait été réduit de plus de la moitié pour atteindre 35,9 %. Et il faudrait un taux de croissance supérieur de 0,2 % seulement (par exemple 2,2 points par an contre 2 points par an) – ce qui ne représente qu'un dixième de l'écart-type transnational – pour obtenir une réduction du taux de pauvreté du même ordre que l'amélioration de la pauvreté d'un écart-type pour une médiane donnée.

Tableau 1
Même de très faibles augmentations de la croissance engendrent une réduction de la pauvreté quasiment identique aux améliorations conséquentes obtenues à un niveau de consommation médiane donné (écart-type du résidu).

Taux de pauvreté	Quartile I de la consommation, seuil de pauvreté à 1,90 $ par jour	Quartile II, 5,50 $ par jour
À la consommation médiane moyenne dans le quartile du pays	72,2 %	74,1 %
Si le taux de pauvreté s'améliore de la valeur d'un écart-type du résidu pour la même consommation	68,6 %	70,2 %
Si la croissance à moyen terme (20 ans) était supérieure de 2 points par an (soit la valeur d'un écart-type transnational des taux de croissance)	35,9 %	51,8 %
Si la croissance à moyen terme (20 ans) augmentait de 0,2 point par an (soit un dixième d'un écart-type transnational des taux de croissance)	67,8 %	72,2 %

Source : calculs de Lant Pritchett à partir des régressions représentées sur la fig. 3.

Cette corrélation extrêmement étroite entre les taux de pauvreté mesurés et le revenu médian/la consommation médiane reste valable en tenant compte des variations dans le temps au sein des pays (KRAAY, 2006)[9]. La fig. 4 indique

9.. L'ensemble de l'étude empirique présentée ici repose sur les sources standards de la Banque mondiale relatives aux revenus/à la consommation des ménages, et non sur des estimations du PIB par habitant. PINKOVSKIY et SALA-I-MARTIN (2016) prétendent, sur la base de données satellitaires sur les éclairages nocturnes, que le PIB par habitant est un indicateur plus performant et plus fiable du progrès, et montre un développement plus rapide et une réduction plus importante de la pauvreté.

un coefficient de détermination R2 de 0,93 entre la variation de la pauvreté au niveau d'« un dollar par jour » (1,90 P $) et la variation du taux de pauvreté prédite seulement sur la base de la variation de la médiane, et la forme fonctionnelle estimée pour la plus longue période observée (supérieure à 10 ans) pour chaque pays.

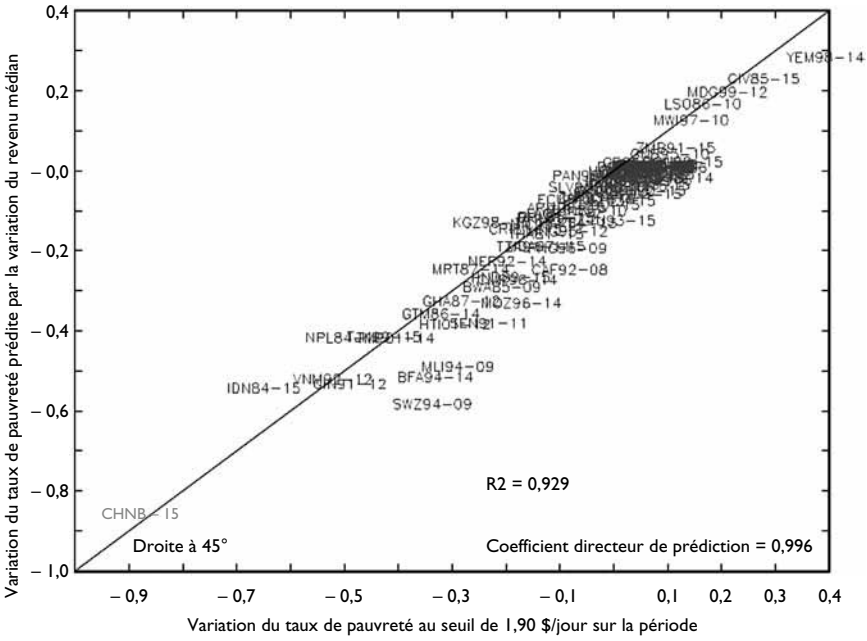


Figure 4

Les variations des taux de pauvreté sont également étroitement liées aux variations du revenu médian/de la consommation médiane.

Source : calculs réalisés par Lant Pritchett avec des données issues de PovcalNet ; l'outil de calcul en ligne pour la mesure de la pauvreté a été conçu par le Development Research Group de la Banque mondiale (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

La fig. 5 montre certains grands pays ayant vu leurs taux d'extrême pauvreté diminuer rapidement, passant de niveaux très élevés à des niveaux très faibles : c'est le cas de la Chine, de l'Indonésie, du Vietnam et, dans une moindre mesure, de l'Inde. Ces baisses des niveaux de pauvreté se sont produites juste sous nos yeux, car nous disposons en effet d'assez bonnes enquêtes sur les ménages, qui permettent un suivi de la pauvreté sur la majeure partie (voire l'ensemble) de ces périodes, grâce auxquelles nous avons pu mener des travaux empiriques minutieux afin d'analyser les facteurs déterminants directs de cette réduction : dans quelle mesure pouvons-nous expliquer cette baisse de la pauvreté par des variations de la tendance centrale (moyenne/médiane), par l'évolution générale des inégalités, ou par des variations de la répartition au-dessous du seuil de pauvreté, dues à l'inégalité moyenne et générale susceptible de répondre à des programmes « anti-pauvreté ». D'après les résultats, il n'est pas exagéré de dire

que « la totalité », voire « plus encore que la totalité » de la réduction de la pauvreté constatée dans ces pays s'explique par les variations de la moyenne/médiane. Ce « plus encore que la totalité » est possible dans la mesure où, dans de nombreux cas, les inégalités se sont aggravées (de façon considérable même, dans le cas de la Chine), l'augmentation de la tendance centrale ayant par conséquent dû compenser cette hausse des inégalités néfaste pour la pauvreté afin de réduire celle-ci.

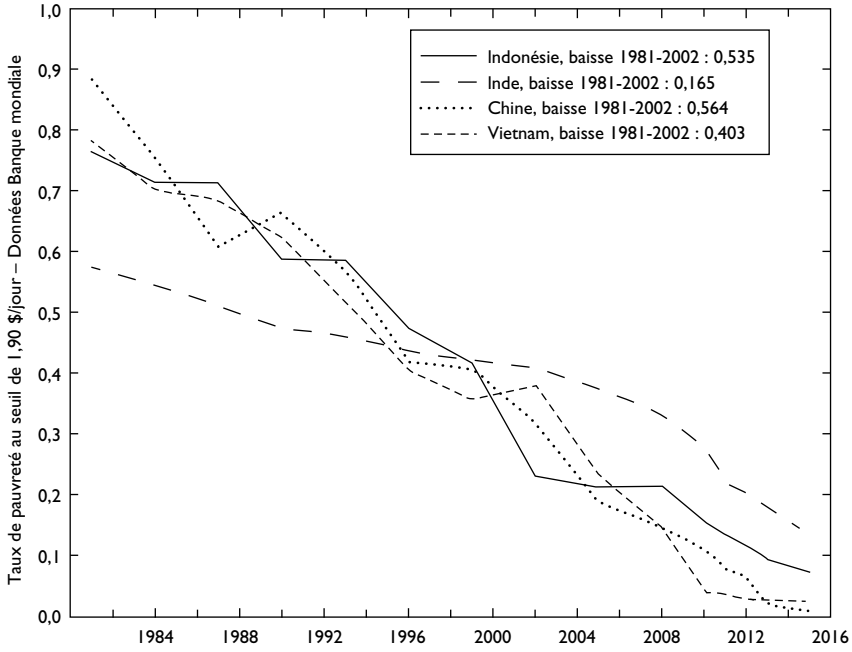


Figure 5

Au début des années 2000, plusieurs pays enregistraient depuis 20 ans les réductions des taux d'extrême pauvreté les plus rapides de l'histoire.

Source : calculs réalisés par Lant Pritchett avec des données issues de PovcalNet ; l'outil de calcul en ligne pour la mesure de la pauvreté a été conçu par le Development Research Group de la Banque mondiale (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx).

Nombreux sont ceux qui semblent penser qu'un « programme anti-pauvreté » est une « intervention » qui fait monter la consommation des « pauvres », à un niveau donné de la médiane. C'est l'effet que l'on attendrait d'un transfert monétaire (conditionnel ou non), d'un programme de « progression » (comme celui précédemment mentionné, encourageant l'élevage d'animaux), de la microfinance, de la fourniture de poulets, des formations à la gestion ou d'à peu près tous les autres programmes anti-pauvreté ciblés, qui sont tous destinés à élever la « queue gauche » de la distribution de la consommation (qui profite aux « pauvres »), tout en maintenant sa tendance centrale à un niveau fixe (voire plus bas, selon la façon dont les programmes sont financés). Les corrélations simples

montrent que les différences de l'incidence des « programmes anti-pauvreté » entre les paires « pays/année », conditionnelle à la médiane, représentent *tout au plus* 1,2 % de la variation transnationale totale des taux de pauvreté[10]. Il s'agit d'une limite supérieure, car, outre la médiane, tous les paramètres – erreur de mesure, différences non liées aux programmes entre la consommation en queue gauche et la consommation médiane (comme les variations des prix relatifs des produits que les pauvres consomment beaucoup), différences de revenus non liées aux programmes et induites par les variations des prix relatifs des actifs détenus par les pauvres (par exemple main-d'œuvre non qualifiée), etc. – totalisent 1,2 % de la variance de la pauvreté observée, de telle sorte que les programmes anti-pauvreté pourraient même ne représenter que 0,1 % (étant donné qu'il existe des programmes adaptés et efficaces dans quelques endroits au moins, il est peu probable que l'on atteigne exactement zéro).

Le développement national et les indicateurs plus larges du progrès social

Outre son impact sur un objectif étroit comme l'extrême pauvreté, le fait d'arriver à des degrés élevés de développement national constitue également une condition nécessaire et suffisante pour atteindre des niveaux importants de bien-être humain global. Il existe une corrélation extrêmement étroite (0,967) entre un indicateur global du bien-être humain (SPI) et le développement national (Pritchett, 2016).

Le SPI[11] est le fruit de la démarche de l'organisation The Social Progress Imperative, visant à mettre en œuvre une nouvelle méthode plus efficiente pour comparer la performance des pays en matière de développement. Ses experts n'utilisent *pas* explicitement le PIB par habitant (ou d'autres indicateurs du développement national), mais se concentrent plutôt sur des mesures directes du bien-être humain. Le SPI comporte trois composants agrégés appelés : (1) besoins humains de base, (2) fondements du bien-être et (3) opportunité. Chacun de ces trois composants repose sur quatre sous-indicateurs, qui sont eux-mêmes construits à partir de mesures spécifiques. L'agrégat « besoins humains de base » (I) comprend par exemple quatre sous-composants (I.1 : « nutrition et soins médicaux de base » ; I.2 : « eau et assainissement » ; I.3 : « logement » ; I.4 : « sécurité des personnes »). Chacun d'eux s'appuie sur des indicateurs spécifiques. Par exemple, le sous-composant I.2 comprend I.2.a : « accès à l'eau courante » ; I.2.b : « accès à une source d'eau assainie en milieu rural » ; et I.2.c : « accès à un système d'assainissement amélioré ». Je ne suis pas en

10. Les mesures et médianes standards de la pauvreté ne sont que des statistiques sommaires différentes de la *même* distribution. L'indice de pauvreté standard (en pourcentage de la population) n'est qu'une intégrale partielle de la distribution au-dessous d'un seuil de pauvreté (j'ai publié des articles sur les méthodes de calcul de la pauvreté, par exemple dans Pradhan *et al.*, 2001). Cela n'implique pas l'existence d'une forte corrélation, car il serait en théorie possible que des programmes puissent « dévier » la distribution et réduire la pauvreté pour une médiane donnée.

11. http://www.socialprogressimperative.org/global-index/

train de dire que le SPI constitue le meilleur indicateur du bien-être humain à l'échelle nationale, mais il s'agit d'une démarche réfléchie et prudente pour mesurer le progrès social d'un pays à un autre. Il utilise 53 indicateurs distincts, dont des indicateurs relatifs à l'économie, à l'éducation et à la santé, mais aussi des indicateurs non standards, comme la tolérance religieuse, l'absence de criminalité et les droits politiques.

J'ai effectué une régression du SPI – qui a été rééchelonné de 0 (plus mauvais score) à 100 (meilleur score) – sur trois indicateurs du développement national : le PIB par habitant (ln) (indicateur de la performance économique), la mesure d'autocratie/démocratie Polity2 (indicateur de la réactivité du système politique) et l'indicateur mondial de gouvernance de l'efficacité du gouvernement (indicateur de la capacité de l'administration), qui attribuent également tous des notes de 0 à 100[12] pour 140 pays (en excluant les pays producteurs de pétrole à hauts revenus et un pays, le Salvador, pour lequel les données du PIB par habitant semblaient fausses). L'indice de développement national (*National Development Index* – NDI) agrège ces trois composants qui sont pondérés à l'aide de coefficients MCO (moindres carrés ordinaires).

La fig. 6 montre que le développement national est empiriquement nécessaire et suffisant pour atteindre des niveaux élevés de SPI. Aucun pays n'atteint un SPI situé dans le tiers supérieur (au-dessus de 70,1) sans présenter un NDI supérieur à 68,6 (soit le niveau de l'Argentine)[13]. De manière analogue, aucun pays situé dans le tiers supérieur du NDI ne présente un SPI inférieur à 61,6.

Le SPI et le NDI ont un coefficient de corrélation de 0,967 (le coefficient de détermination R2 de la régression étant de 0,935). Il s'agit d'une relation étonnamment étroite entre deux indicateurs différents, tant sur le plan conceptuel qu'empirique, si l'on considère que des mesures transnationales différentes d'une même chose, réalisées à partir de sources ou de méthodes différentes – comme les « années de scolarité de la population adulte » ou la « mortalité infantile » – ne présentent bien souvent pas des taux de corrélation transnationaux de 0,96, ne serait-ce que du fait des simples erreurs de mesure.

12. Je ne pense pas que cela dépende de l'utilisation de ces trois indicateurs spécifiques du développement national. L'indice des états fragiles (FSI) produit par le Fund for Peace, par exemple, comprend de multiples composants, dont deux, les « services publics » et la « légitimité de l'État », peuvent être utilisés comme indicateurs empiriques alternatifs pour les concepts de « capacité administrative » et de « réactivité politique ». Une régression de l'indice de progrès social global sur le PIB par habitant, le composant « services publics » du FSI et le composant « légitimité de l'État » du FSI (tous dimensionnés sur une échelle de 100) donne un coefficient de détermination R2 de 0,947 (ou plus), ces trois indicateurs jouant un rôle fort.

13. Des mesures du bien-être humain tendent parfois à souligner que le PIB par habitant est un indicateur peu fiable du bien-être humain (pour lequel aucun économiste ne le propose évidemment jamais) en montrant des « cas particuliers » qui atteignent des SPI élevés avec des PIB par habitant (relativement) bas. Le « développement national » *inclut* toutefois aussi les critères de politique, de capacité de l'État et de transformation sociale. Et si l'on prend en compte cette définition plus large, des pays qui peuvent parfois s'avérer très performants en termes de PIB par habitant, comme le Costa Rica (sur la figure, CRI se superpose à URY, c'est-à-dire l'Uruguay), ont un SPI élevé et « sur-performant » même en matière de NDI, sans être considérés comme des « cas particuliers aberrants », étant donné qu'ils présentent un NDI élevé.

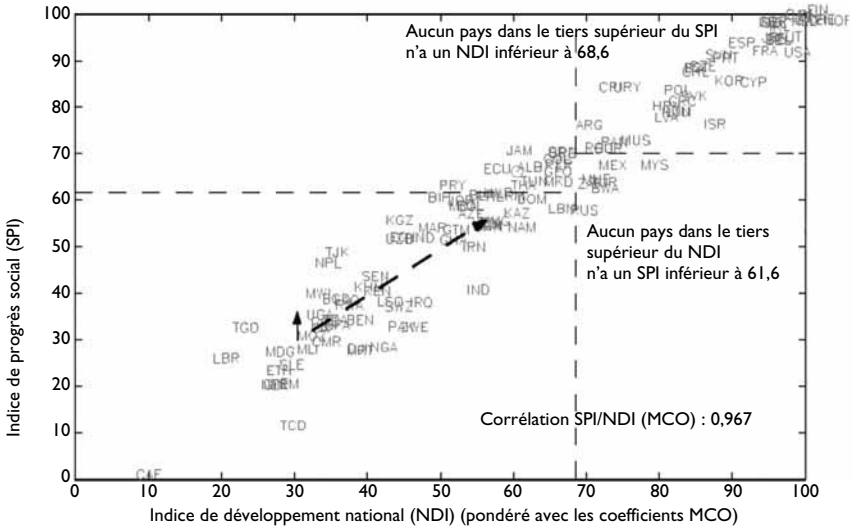


Figure 6

Le développement national est empiriquement nécessaire et suffisant pour atteindre des niveaux élevés de l'indice de progrès social.

Source : calculs de Lant Pritchett à partir des données et méthodes décrites dans le texte.

Comme pour la pauvreté, cette relation forte et étroite implique que les gains potentiels en termes de progrès social pour un niveau de développement national donné sont relativement faibles (par rapport à l'amplitude du SPI). Le Mozambique (abréviation MOZ) a approximativement le même SPI réel et estimé (et par conséquent le même NDI) d'environ 30. Supposons que, d'une manière ou d'une autre, le Mozambique soit « ultra performant » en matière de progrès social pour un niveau de développement national donné, dans le sens spécifique où son SPI serait supérieur de la valeur d'un écart-type (et que, dans l'hypothèse d'une distribution normale, il se trouve donc, avec son NDI, dans le 84e centile des pays plutôt que dans le 50e). Son SPI serait alors de 36 (effet illustré par la flèche verticale sur la fig. 6). Même si ce gain n'est pas neutre, le SPI du Mozambique resterait inférieur à celui du Laos, du Bangladesh ou du Kenya. En revanche, si le Mozambique améliorait son résultat de la valeur d'un écart-type au niveau de chacun des éléments du développement national, son SPI atteindrait 56, et serait ainsi supérieur à celui des pays à revenu intermédiaire de la tranche supérieure comme le Maroc ou l'Indonésie (flèche pointillée pointant vers le « nord-est » sur la fig. 6).

Le tabl 2 montre la relation empirique entre les trois composants et les 12 sous-composants du SPI et les indicateurs du développement national. Chacun des trois composants du SPI a une corrélation très étroite avec le NDI (« besoins de base » : 0,904 ; « Fondements du bien-être » : 0,925 et « Opportunité » : 0,932). Les 12 sous-composants (sauf un[14]) sont eux aussi fortement liés au développement national.

14. L'indicateur pour lequel il n'existe pas de corrélation positive forte est la « qualité de l'environnement », qui inclut les émissions de gaz à effet de serre, ce sous-composant étant positivement associé au PIB par habitant.

Tableau 2

L'indice de progrès social, ainsi que tous ses composants et sous-composants sont fortement corrélés aux trois indicateurs du développement national.

Indicateur du progrès social, ses 3 composants (besoins humains de base, fondements du bien-être et opportunité), et les 4 sous-composants de chaque composant	**Productivité économique ([ln] PIB par habitant, Penn World Tables (PWT) 8.0, rééchelonné de 0 à 100)**		**Capacité administrative (indicateurs mondiaux de gouvernance, efficacité du gouvernement, rééchelonnés de 0 à 100)**		**Réactivité politique (projet Polity IV, indicateur Polity 2, rééchelonné de 0 à 100)**		**Coef. de détermination de la régression sur les indicateurs du développement national**
	Coefficient MCO	**t-stat.**	**Coefficient MCO**	**t-stat**	**Coefficient MCO**	**t-stat**	
Indice de progrès social	0,53	13,67	0,34	7,38	0,12	5,01	0,935
I) Besoins humains de base	0,74	12,10	0,18	2,46	- 0,02	- 0,43	0,835
I.1) Alimentation et soins médicaux de base	0,57	8,86	0,34	5,17	0,18	5,06	0,865
I.2) Eau et assainissement	0,31	4,95	0,51	8,15	0,23	7,11	0,873
I.3) Logement	0,80	9,74	- 0,09	- 0,95	0,04	0,79	0,672
I.4) Sécurité des personnes	1,17	11,78	0,01	0,06	0,06	1,12	0,784
II) Fondements du bien-être	1,06	13,30	0,04	0,47	- 0,01	- 0,36	0,820
II.1) Accès aux connaissances de base	- 0,02	- 0,27	0,77	7,86	- 0,09	- 1,83	0,603
II.2) Accès à l'information et aux moyens de communication	1,00	10,62	- 0,11	- 1,09	0,04	0,73	0,707
II.3) Santé et bien-être	0,53	8,02	0,22	3,25	0,21	6,11	0,816
II.4) Qualité de l'environnement	- 0,18	- 1,55	0,50	4,34	0,01	0,13	0,242
III) Opportunité	0,11	1,33	0,52	6,43	0,18	4,34	0,709
III.1) Droits individuels	- 0,08	- 0,86	0,53	5,68	0,55	11,58	0,765
III.2) Liberté individuelle et choix	0,16	2,06	0,66	8,65	- 0,01	- 0,37	0,757
III.3) Tolérance et inclusion	0,19	1,71	0,41	3,70	0,14	2,48	0,517
III.4) Accès à l'enseignement supérieur	0,93	11,21	0,17	2,04	0,03	0,73	0,824

Source : calculs de Lant Pritchett.

Les indicateurs nationaux relatifs au bien-être évalué subjectivement sont également fortement corrélés avec le développement national. La régression de « l'échelle de vie » de Cantril, qui mesure le bien-être subjectif moyen, sur les trois indicateurs du développement national donne un coefficient de détermination R2 de 0,66 (coefficient de corrélation de 0,812 avec un NDI pondéré par la méthode MCO). Le *World Happiness Report* a développé un autre indice du bien-être humain basé sur le lien empirique entre sept facteurs (tels que les « perceptions de la corruption », l'« espérance de vie en bonne santé », l'« aide sociale », ainsi que des mesures de l'état affectif) et l'« échelle de vie » qui mesure le bien-être subjectif. Un indice composé de ces six éléments de l'indice du bonheur à pondération égale et régressé sur les trois indicateurs du développement national, donne un coefficient de détermination R2 de 0,788 sur 120 pays (coefficient de corrélation de 0,887 avec un NDI pondéré par la méthode MCO). Là encore, la corrélation entre cet indice du « bonheur » à six éléments et la mesure de « l'échelle de satisfaction de vie » observée directement est de 0,81. Si ces résultats sont inférieurs à la corrélation SPI/NDI, les trois indicateurs du bien-être humain (le SPI, la satisfaction de vie subjective et le *World Happiness Report*) sont presque aussi étroitement corrélés entre eux que chacun ne l'est avec un indice de développement national (spécifique à la mesure).

Le développement national permet d'éradiquer la pauvreté et d'atteindre des niveaux élevés de bien-être humain

Grâce au recueil de données plus nombreuses et de meilleure qualité, nous pouvons montrer que les rapports entre le développement national et la pauvreté, le bien-être humain global, ou des indicateurs spécifiques du bien-être sont aussi forts et étroits que cela a été affirmé.

Ce qui est curieux, c'est que l'on ait pu en douter. Le développement national à quatre dimensions est une « machine à bien-être humain ». Considérons n'importe quel objectif susceptible d'apporter un niveau de bien-être important et largement répandu, comme l'accès à l'eau, une meilleure santé, un logement amélioré, une plus forte scolarisation : le développement national est justement conçu pour favoriser la réalisation de cet objectif. Une économie plus productive, qui génère une augmentation générale des revenus offre aux ménages des moyens plus importants pour poursuivre leurs objectifs. Et si ces objectifs se rapportent à des biens privés, il paraîtrait bien étrange qu'un accroissement des revenus privés n'entraîne pas une hausse des niveaux de consommation – et, d'ailleurs, de tout ce qui importe empiriquement dans les composants du SPI liés à « l'eau et à l'assainissement », au « logement » et à l'« accès aux connaissances de base », le seul corrélat significatif était le PIB par habitant[15].

15. Et on s'attendrait à ce que le rapport avec le développement national soit plus fort/étroit encore pour les « produits de première nécessité », les économistes définissant un « produit de première nécessité » comme un bien dont l'utilité marginale devient très forte lorsque sa consommation chute et, par association, comme un bien dont l'élasticité-prix est censée être très faible (en particulier à des niveaux très bas). La simple loi d'Engel selon laquelle la part des produits alimentaires dans la consommation décroît de manière linéaire avec l'augmentation du revenu/de la consommation est incontestablement le phénomène le mieux documenté de toute l'économie.

Mais, lorsque les objectifs du bien-être humains font intervenir des « biens publics » (non rivaux et non exclusifs), ou que les marchés pour ces biens sont « défaillants », on est alors précisément en présence de problématiques auxquelles des gouvernements réactifs et efficients peuvent remédier. Et en effet, pour le composant « qualité de l'environnement », les seuls corrélats forts partiels sont la capacité et le système politique, et non le PIB par habitant, de même que le seul corrélat partiel pour la « sécurité des personnes » est la capacité de l'État. Personne, même l'économiste le plus fervent et le plus orienté « marché », n'a jamais démontré que le revenu pourrait à lui seul résoudre tous les problèmes. La réactivité et la capacité de l'État ont toujours été des éléments incontournables de la vision du développement.

Les RCT en développement en tant que méthode pour améliorer le bien-être humain

Revenons à la « folie ». Comment en sommes-nous arrivés aux études sur les poulets ? Comment l'économie du développement en est-elle arrivée à penser à tout sauf au développement national ? Comment quelqu'un pourrait-il argumenter, justifier ou cautionner l'affirmation qu'une étude de l'efficacité relative de programmes ciblés de type « fourniture de volailles » *versus* « transferts monétaires » mettant en œuvre une méthode particulière peut être le « meilleur investissement » pour lutter contre la pauvreté ? Cette prétention comporte trois facteurs multiplicatifs : (a) la probabilité qu'une étude fournisse des connaissances fiables et exploitables, (b) la probabilité que ces connaissances changent les événements qui se produisent dans le monde d'une façon susceptible d'améliorer la situation, et (c) le gain total en termes de bien-être humain (en évaluation normative) résultant de tels changements (fig. 7).

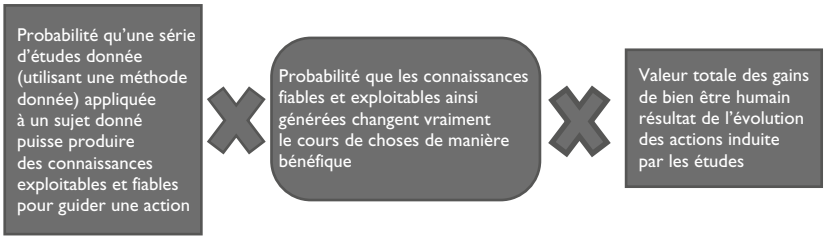


Figure 7
Grandeurs empiriques à déterminer pour prendre des décisions sur la valeur relative attendue de différents types d'investissements pour la recherche.
Source : Lant Pritchett.

La fig. 8 représente de façon descendante la cartographie des liens entre les mesures du bien-être humain (indicateurs ou domaines composés/agrégés et spécifiques), et montre si une évaluation normative les classe comme « fétichistes » ou non. Le principe de la mesure dite « fétichiste » ne repose *pas* sur le fait que les « plus pauvres » (c'est-à-dire les personnes disposant d'un accès moindre à l'hygiène/l'éducation/l'énergie) bénéficient d'une pondération plus élevée dans l'indicateur de bien-être, et les « plus riches » (c'est-à-dire ceux qui possèdent une quantité supérieure d'un élément particulier) d'une pondération moins importante. N'importe quel indicateur standard d'inégalités, comme l'indice d'Atkinson ou une fonction de bien-être social (*Social Welfare Function* – SWF) standard basée sur l'hypothèse d'une utilité marginale décroissante peut prendre en compte cet aspect (en utilisant des paramètres qui confèrent divers degrés de « préférence pour les pauvres »), ou de manière similaire des mesures sectorielles peuvent donner davantage de poids à des niveaux de service spécifiques ou à certains groupes. L'essence même d'une mesure « fétichiste » est que le gain de bien-être humain au-dessus d'un seuil arbitraire, quel qu'il soit (un seuil de pauvreté, l'« achèvement du cycle d'enseignement primaire » ou « l'accès à des latrines ») est *exactement égal à zéro.*

Sur la fig. 8, du bas vers le haut, les flèches illustrent les revendications sur l'intensité/ampleur des impacts causaux du développement national (*National Development* – ND), des programmes ciblés, liés aux revenus (*Targeted programs* [*Income*] – TP [Y]) ou à des indicateurs sectoriels spécifiques (*Targeted Programs* [*Sectors*] – TP [S]), ou des réformes sectorielles (*Sector-Wide Reforms|National Development* – SR|ND) sur le bien-être humain.

L'allégation de « meilleur investissement » consiste à dire que le lien entre l'étude RCT et le programme ciblé amélioré visant à augmenter les revenus TP (Y) (flèche en tirets et en points) fois les gains générés par TP (Y) sur le développement « fétichiste » (« extrême pauvreté » ; flèche plus petite) est supérieur, en termes de rapport coûts-bénéfices, à n'importe quel autre. L'allégation opposée, selon laquelle la recherche sur le développement national est meilleure, est fondée sur l'une ou l'autre des deux revendications suivantes :

– (a) l'impact d'une recherche non basée sur les RCT sur les résultats obtenus au niveau du développement national (flèche à tirets) multiplié par l'impact du développement national sur un indicateur de développement « fétichiste » (par exemple l'« extrême pauvreté ») (grosse flèche) est supérieur ;

– (b) l'impact d'une recherche non basée sur les RCT sur les résultats obtenus au niveau du développement national (flèche à tirets) multiplié par l'impact des résultats du développement national sur le bien-être humain global (agrégé) (ajusté des inégalités) est supérieur en valeur (quel que soit le système de valorisation raisonnable) à celui d'une RCT portant sur TP (Y) sur l'extrême pauvreté.

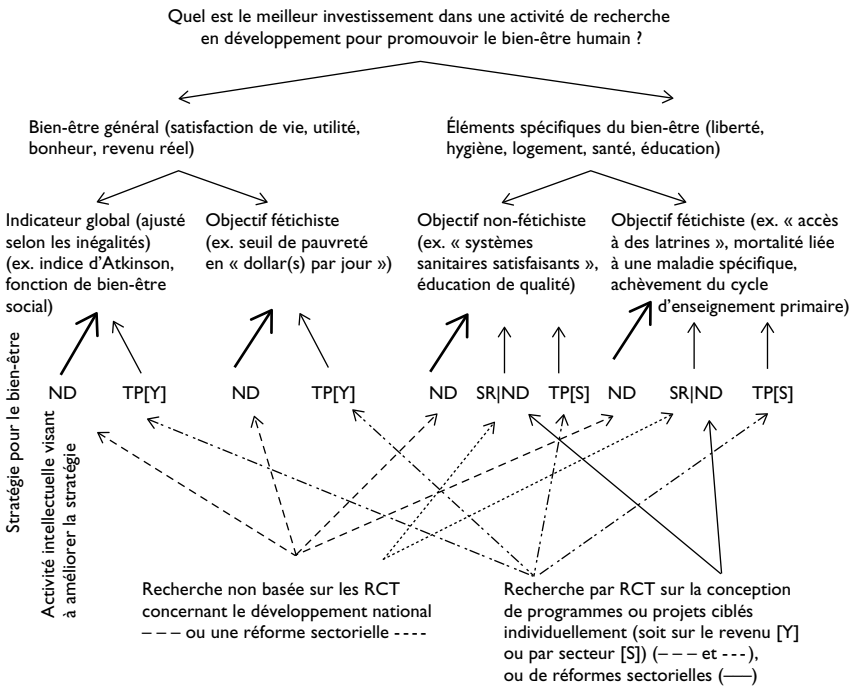


Figure 8

Quel est le meilleur investissement pour l'activité de recherche en développement en matière de promotion du bien-être humain ?

Source : Lant Pritchett.

Presque tous les économistes s'accordent sur deux points dans les fig. 7 et 8. Premièrement, ils s'entendent sur le fait que l'ampleur, en dollars, du gain généré par le développement national et les réformes sectorielles est considérablement plus grande que celle du gain susceptible d'être obtenu avec des programmes ciblés. Les *randomistas* ne soutiennent généralement pas que les bénéfices de la croissance sur la pauvreté ne seraient pas importants, étant donné que, d'après les chiffres mentionnés plus hauts, cette affirmation serait manifestement erronée, mais plutôt que l'impact de la recherche sur la croissance serait très faible, voire nul.

Deuxièmement, ils sont d'accord sur le fait que l'impact des RCT sur le développement national ou les réformes sectorielles est très certainement limité. L'une des raisons pour lesquelles j'ai souligné que les processus de développement national s'opèrent à un niveau ontologique supérieur à l'individu est que les RCT ne sont généralement possibles (et ne peuvent atteindre une puissance statistique suffisante) que lorsqu'un nombre important d'unités peuvent être soumises aux statuts de « traitement » et de « contrôle », ce qui est impossible pour les phénomènes qui ont lieu à l'échelle d'une économie, d'une politique nationale ou d'une organisation.

Les allégations les plus courantes émises par les fervents défenseurs des RCT sont en conséquence soutenues par plusieurs arguments. D'abord, si l'impact du développement national sur tous les types d'indicateurs du bien-être est important, et que le développement national est à la fois suffisant pour atteindre des objectifs étroits et nécessaire pour atteindre des objectifs ambitieux, l'impact de la recherche sur le développement national est quant à lui quasi nul (les flèches représentées par des tirets entre les recherches non liées aux RCT et le développement national n'existent généralement pas). Donc, même si l'impact des RCT sur les programmes ciblés (liés aux revenus Y ou à des indicateurs spécifiques S) est faible, et ne s'applique qu'à des objectifs étroits (« fétichistes »), la recherche est rentable, ne serait-ce que parce qu'elle a un effet, alors que les autres types de recherche ont une efficacité (quasi) nulle. Ensuite, une autre argumentation consiste à dire que la mesure du bien-être humain est exclusivement appréhendée par des indicateurs « fétichistes », et que les gains au-delà du seuil n'ont pas d'importance, donc que l'impact du développement national sur les objectifs « non fétichistes » n'ont de ce fait que peu de valeur.

Que l'on considère l'une ou l'autre des revendications exposées ci-dessus, il convient d'ajouter *que* :

– les RCT doivent générer des connaissances fiables et utiles sur les programmes ciblés destinés à augmenter les revenus ou à atteindre des objectifs spécifiques ;

– les connaissances fiables et utiles produites par les RCT doivent changer réellement le cours des événements, c'est-à-dire qu'elles doivent permettre de lever des contraintes majeures à la mise en place de meilleurs programmes ciblés.

Allégation largement admise numéro 1 : l'ampleur des gains issus du développement national est beaucoup plus importante que celle des programmes ciblés

Kenny et Pritchett (2013) montrent que, fondamentalement, pour toutes les mesures du bien-être humain, les effets de la progression du développement national (phénomène appelé « *drive* ») ou les gains d'efficacité sectorielle (appelé « *shift* ») dépassent considérablement les bénéfices tirés des programmes ciblés (appelé « *kink* »).

Pritchett *et al.* (2016) évaluent la valeur actualisée nette (VAN) du PIB ajouté (ou perdu) par rapport à un contrefactuel « *business as usual* » lors de différents épisodes de croissance ou de récession économique. Notre méthodologie consistant à définir les dates et l'amplitude des épisodes de croissance suggère que les périodes d'accélération de la croissance qui ont eu lieu en Chine en 1977 et en 1991 ont généré des gains de VAN (valeur actuelle nette) à hauteur de 2 650 milliards et 11 800 milliards de dollars (soit au total plus de 14 000 milliards). En Inde, les accélérations de la croissance survenues en 1993 et en 2002 ont produit des gains de 1 100 milliards et de 2 500 milliards (soit au total 3 600 milliards). L'accélération de la croissance enregistrée par l'Indonésie en 1967 a conduit à un gain de VAN de 1 100 milliards par rapport

au contrefactuel de référence. Si les gains absolus résultant de l'accélération de la croissance du Vietnam en 1989 étaient inférieurs, à hauteur de 455 milliards de dollars, ils représentaient toutefois un gain de VAN de 6 911 $ par habitant. Ces épisodes de croissance ont en outre été accompagnés d'une baisse rapide de l'« extrême pauvreté », celle-ci atteignant des niveaux très faibles (fig. 5). De manière analogue, les décélérations produisent des pertes tout aussi importantes par rapport au taux de croissance « *business as usual* ». L'épisode de décélération rencontré par le Brésil en 1980 s'est ainsi chiffré par des pertes à hauteur de 7 500 milliards de dollars, alors que celles subies par l'Indonésie suite à la crise qui a frappé l'Asie de l'Est en 1996 ont approché les 1 000 milliards de dollars, et qu'au Mexique les pertes combinées induites par les décélérations de 1981 et de 1989 ont atteint les 1 500 milliards. Plusieurs pays africains ont aussi connu des décélérations de la croissance qui se sont traduites par des pertes certes réduites en valeur absolue, mais importantes en termes de VAN par habitant, à savoir : 9 600 $ par habitant au Malawi en 1978, 13 300 $ au Kenya en 1967 et 15 200 $ en Côte d'Ivoire en 1978.

L'expérimentation sur « les animaux d'élevage » publié dans *Science* a montré qu'un programme de « progression » complexe et multidimensionnel destiné à des foyers en situation d'extrême pauvreté avait augmenté leurs revenus en année 3 dans 5 des 6 sites concernés par l'étude. À titre d'indication des ordres de grandeur, sur la base de la moyenne des cinq sites, des dépenses de 4 545 $ par foyer engagées en années 1 et 2, ont généré un gain de 344 $ par foyer ou, en partant de l'hypothèse qu'un foyer type compte 4 personnes, de 86 $ par personne. En supposant que ce montant obtenu en année 3 perdure dans le temps, on obtient, avec un taux d'actualisation de 5 %, un gain brut moyen de VAN de 8 472 $ par foyer, soit un taux de rendement de 7 % environ. En partant de l'hypothèse de 4 personnes par foyer, on peut dire qu'un investissement de 1 136 $ par personne génère un gain ponctuel de 86 $ en année 4. Supposons maintenant que nous voulions exploiter les données issues de cette évaluation « étalon-or » d'un programme anti-pauvreté pour augmenter les revenus à hauteur d'une VAN de 6 911 $ au Vietnam. Le montant des investissements dans le programme s'élèverait alors à 333 milliards de dollars, soit plus que le PIB total *actuel* (post-croissance) du Vietnam ou environ trois fois l'aide mondiale *totale* au développement.

Les bénéfices que l'on peut tirer de systèmes financiers performants, et surtout de l'évitement d'une crise majeure, sont énormes. Les pertes de PIB des pays de l'OCDE résultant de la crise financière de 2008 étaient estimées à environ 3,5 %, ou 1 900 milliards de dollars en 2014. Si l'on considère qu'il s'agit d'une perte « définitive » par rapport à un contrefactuel hors crise économique, la VAN (avec un taux d'actualisation de 5 %) atteint 38 200 milliards de dollars. La Réserve fédérale des États-Unis évalue la VAN de la perte pour le pays à 70 000 $ par habitant. En 2016, le stock d'actifs total de la microfinance s'élevait à environ 102 milliards de dollars. Supposons, dans la perspective la plus positive et la plus folle qui soit, que le gain annuel pour les emprunteurs soit de 10 % des actifs, celui-ci serait de 10,2 milliards de dollars. Présumons

aussi, là encore avec un optimisme extrême, qu'une étude rigoureuse puisse, d'une façon ou d'une autre, permettre de *doubler* ce gain (par rapport à un contrefactuel), les clients de la microfinance gagneraient alors 10,2 milliards de dollars supplémentaires au total. Les pertes engendrées par une seule crise financière mondiale (de grande portée) seraient 200 fois supérieures aux gains tirés du doublement des plus-values totales de la microfinance.

Il est primordial d'élever les niveaux d'apprentissage dans l'enseignement de base donné aux enfants afin de les préparer à vivre au XXIe siècle. Au regard du défi que cela représente, quelle est l'importance de la recherche sur les impacts des transferts monétaires conditionnels en matière de scolarisation ? Sur la base d'une évaluation récente de l'enseignement en Zambie dans le cadre du programme international pour le suivi des acquis des élèves (Pisa) pour le développement (Pisa-D), j'ai estimé que, sur les 360 000 enfants âgés de 15 ans en Zambie, seulement 36 % allaient à l'école et faisaient l'objet d'une évaluation, et que, parmi eux, un total estimé à cinq enfants seulement (je ne parle pas de 5 %, mais de cinq élèves, comme les cinq doigts de la main) avait un niveau de compétence global élevé en lecture (soit les niveaux 4 ou plus de Pisa, atteints par environ un tiers des élèves de l'OCDE). Même si, par on ne sait quel effort héroïque dont, par exemple, le recours étendu aux transferts monétaires conditionnels, le taux de scolarisation des enfants de 15 ans atteignait 100 %, compte tenu des niveaux d'apprentissage actuels, on compterait seulement *14 élèves* supplémentaires capables de maîtriser la lecture à des niveaux de compétences élevés. Or, le Vietnam est considérablement plus performant que la Zambie en matière d'enseignement, à des degrés qui ne relèvent pas de programmes ciblés, mais semblent plutôt être le fruit du bien meilleur fonctionnement du système éducatif dans son ensemble.

Allégation largement admise numéro 2 : les études RCT ne répondent pas à la problématique du développement national

PRITCHETT (2014a) s'appuie sur l'article dans lequel VIVALT (2020) passe en revue les résultats de RCT pour comparer les sujets sur lesquels un nombre suffisant de RCT a été réalisé pour permettre de déterminer, par des questions simples, si un sujet X peut être un facteur majeur de croissance. Et aucun des domaines courants couverts par des RCT (transferts monétaires conditionnels, microfinance, fours de cuisson améliorés, vermifugation) ne constitue un déterminant important crédible du niveau de revenu ou de croissance. Leurs partisans ne le prétendent d'ailleurs pas. La raison pour laquelle j'ai insisté sur le caractère social de la transformation dans quatre dimensions du développement national est que la réussite d'une RCT en tant que stratégie de recherche passe nécessairement par (a) une affectation (raisonnablement) irréprochable des unités dans les groupes de « traitement » et de « contrôle », et (b) un nombre suffisant d'unités pour assurer une puissance statistique suffisante. Ceci explique presque inévitablement pourquoi cette méthode convient bien aux actions individualisables

(ou à l'échelle de petites unités, comme une clinique, une école ou un poste de police) et n'est en revanche pas adaptée pour étudier l'impact d'une politique sur la performance d'un marché, l'évolution de la gouvernance d'un État ou la transformation sociale. Même si une RCT devait traiter de ces questions (une étude sur l'information et le comportement des électeurs, par exemple), elle le ferait d'une façon qui, si les résultats étaient extrapolés à l'échelle correspondante, n'aurait pas plus de « rigueur » et de valeur comme preuve que n'importe quelle autre méthode, car, pour pouvoir utiliser la méthode avec précision, les effets d'« équilibre général » à l'échelle de l'ensemble du système devraient être mis entre parenthèses.

Allégation nécessaire, mais fausse numéro 1 : l'impact de toute recherche (RCT ou autre) sur le développement national (ou des réformes sectorielles) est infime

Étant donné l'ampleur relative des gains de bien-être humain issus du développement national et l'inadéquation de la méthode RCT à la promotion du développement national ou des changements sectoriels, le seul argument possible pour les partisans des RCT est que le développement national, qui inclut la croissance économique, est globalement impénétrable et résiste à tout type de recherche.

Cet argument est contredit par des interprétations communément admises d'événements se déroulant dans bon nombre de pays. D'une part, certains pays (comme la Chine, l'Inde, le Vietnam, l'Indonésie) ont affirmé : « sur la base de notre interprétation des preuves existantes (dont celles émanant des économistes), nous allons passer de la politique X à la politique Y afin d'accélérer notre croissance » ; ces pays sont effectivement passés de la politique X à la politique Y, et (3) ont bien enregistré une accélération importante (voire très importante) de leur croissance par rapport à la base de référence mesurée par des méthodes standards (Pritchett *et al.*, 2016). Il faudrait être particulièrement têtu et habile pour faire valoir avec succès que : « Les responsables politiques ont changé de politique pour favoriser la croissance en s'appuyant sur les preuves avancées et [que] la croissance s'est accélérée, mais (a) c'était juste un coup de chance ; en réalité, ce n'est pas ce changement de politique qui a engendré l'évolution de la croissance, c'est autre chose, ou (b) (remarque plus subtile) les politiques adoptées ont fonctionné, mais c'était juste un coup de chance, car il n'y avait pas suffisamment de preuves que ce serait bien le cas pour que l'on puisse considérer que ce sont bien les "preuves" qui ont entraîné ce changement de politique ».

Il y a aussi un assez grand nombre de pays qui ont fait l'inverse. Des économistes (nationaux ou étrangers) ont dit aux dirigeants : (1) « si vous persistez à appliquer la politique X, vous allez subir de lourdes (voire très lourdes) conséquences négatives pour la croissance économique de votre pays » ; (2) ces dirigeants ne les ont pas écoutés ; (3) les conséquences négatives annoncées se sont bel et bien concrétisées. En 2018, l'économie vénézuélienne ne s'est pas retrouvée en situation d'hyperinflation et de dépression parce que « les économistes ont peu

de choses utiles à dire sur la croissance économique », dans le sens où, si leur avis avait été suivi, il n'aurait pas été utile. Si l'argument consiste à dire que les études peuvent fournir une orientation fiable, mais que cela ne signifie pas pour autant que ce conseil va transformer le cours des événements, alors la question n'est pas de savoir si cela fonctionne dans tous les cas, mais si cela ne contribue *jamais* à changer le cours des événements. Il existe également des exemples de gouvernements qui avaient dit « sur la base de ce que les économistes affirment, nous allons changer de politique pour éviter une récession massive/hyperinflation » et qui l'ont fait, et cela a marché (du moins dans le sens où aucune crise ne s'est produite). Alors que les « accélérations de la croissance » ont pu être difficiles à anticiper avec des politiques standards (HAUSMANN *et al.*, 2005), des preuves empiriques montrent que les « effondrements de la croissance » sont un peu plus prévisibles (BREUER et MCDERMOTT, 2013).

Cela ne veut pas dire que toutes les affirmations sur les politiques de croissance qui s'appuient sur des études ont toujours été justes. Les « décennies perdues » en Amérique latine et la « transition-dépression » subie par certains pays (mais pas tous) autrefois dominés par l'Union soviétique sont deux cas dans lesquels des politiques de croissance ont été adoptées sur la base de recommandations d'économistes, mais semblent ne pas avoir fonctionné. Toutefois, comme le montrent Garchitorena *et al.* (chap. 5, ce volume), parmi les dix médicaments les plus prescrits, beaucoup ne marchent que sur un tiers des patients. Ainsi, ce n'est pas parce qu'une recommandation ne connaît pas une réussite universelle qu'elle n'est pas bonne. Si je peux vous donner une astuce qui augmente de 10 % vos chances de gagner un million de dollars à la loterie, cela en vaut globalement la peine. Des études récentes suggèrent en outre que la position qui consiste à penser que « les recherches sur la croissance sont une plaie » et que leurs recommandations n'ont jamais aucune valeur est trop extrême (par exemple, EASTERLY [2019] sur le consensus de Washington ; IRWIN [2019] sur le commerce).

N'oublions pas les résultats du tabl. 1, qui montrent que les effets escomptés des études sur la croissance, aussi faibles soient-ils, peuvent néanmoins produire une réduction de la pauvreté du même ordre que celle qui peut être attendue de l'amélioration des programmes anti-pauvreté. Supposons que l'on donne des conseils sur la croissance à 10 pays, que 9 d'entre eux n'appliquent pas ces conseils ou les appliquent sans que cela produise d'effets positifs, mais que, dans 1 des 10 pays, la croissance augmente de 2 points par an sur 20 ans. Malgré un tel manque d'efficacité, l'impact de ces conseils en matière de réduction de la pauvreté est aussi important que celui qu'on peut espérer d'une amélioration des programmes anti-pauvreté. Et, évidemment, si les pays qui adoptent ces recommandations sont de grands pays, comme la Chine, l'Inde, le Vietnam ou l'Indonésie dans les années 1960, alors l'augmentation globale du bien-être est énorme, même si les recommandations sont, dans une large mesure, inefficaces.

De plus, la faible efficacité des conseils dispensés sur la croissance dans les années 1980 et 1990 pourrait tout aussi bien conduire à recommander de mener

beaucoup plus (et non pas moins) d'études sur la façon de favoriser le développement national, sachant à quel point il est important d'obtenir de bons (plutôt que de mauvais) conseils sur ces questions absolument fondamentales. Ce n'est pas comme si les économistes versaient dans la complaisance, et choisissaient d'ignorer les impacts négatifs pour la croissance de nombreux épisodes de réformes politiques (par exemple, World Bank, 2005), ou de s'en tenir à des « régressions de la croissance insensées ». Une approche prenant en compte le caractère épisodique de la croissance des pays en développement (par exemple, Ben-David et Papell, 1998 ; Pritchett, 2000 ; Jones et Olken, 2008 ; Berg *et al.*, 2012), combinée à une démarche-diagnostic (par exemple, Hausmann *et al.*, 2008 ; Hausmann *et al.*, 2008 ; Rodrik, 2009) était en train de faire ses preuves alors même que le mouvement des *randomistas* prenait son envol[16].

Allégation nécessaire, mais fausse numéro 2 : l'évaluation du bien-être humain relève du « fétichisme »

L'autre voie illustrée sur les fig. 7 et fig. 8 pour justifier une priorisation des RCT dans l'activité de recherche en matière de développement est de se focaliser exclusivement sur des mesures étroites ou « fétichistes » du bien-être humain. Avec cette approche, les idées que le développement national soit une condition nécessaire pour atteindre des niveaux de bien-être modérés à élevés, et que des gains très importants peuvent être retirés du développement national deviennent moins convaincantes. J'ai expliqué en détail dans d'autres publications les raisons pour lesquelles les objectifs fétichistes de manière générale, et plus spécifiquement ceux liés à l'extrême pauvreté, sont illégitimes à tout point de vue, que ce soit sur le plan économique (Pritchett, 2006 ; 2013a), moral (Pritchett, 2014c), politique (Gelbach et Pritchett, 2002 ; Pritchett, 2005 ; 2014a ; 2014b) en tant qu'objectifs pour le développement (Pritchett, 2015), ou pour les organisations de développement (Pritchett, 2013a), donc je serai bref. L'argument simple, mais non moins irréfutable, à opposer aux objectifs fétichistes, tant en matière de revenus que d'indicateurs spécifiques (comme l'eau ou l'éducation) est fondé sur « l'introspection + la règle d'or ».

L'introspection

La nature même des mesures « fétichistes » est que l'augmentation du bien-être devient *exactement* nulle lorsque l'on dépasse un seuil très bas. Posez-vous la question pour vous-même : est-ce que la valeur que vous accordez à vos revenus est devenue exactement nulle à partir du moment où votre revenu a dépassé un

16. Détournant par la même occasion de la recherche sur la croissance des financements qui lui étaient destinés. À titre d'exemple, l'article Crépon *et al.* (2015), réexaminé par Bédécarrats *et al.*, (2021) et évoqué plus loin (chap. 7, ce volume), a fait l'objet d'un financement et d'une promotion par l'International Growth Centre, qui était initialement financé par le Department for International Development (DFID) pour améliorer « les analyses de la croissance » et donner aux pays des recommandations mieux définies et plus pragmatiques sur la mise en œuvre de politiques destinées à favoriser la croissance. Peu importent les mérites ou démérites de fond de cette publication, elle concerne bien un programme ciblé, et personne ne peut prétendre qu'elle vise à promouvoir le développement national ou la croissance.

seuil relativement bas ? Est-ce que votre volonté de payer pour des installations sanitaires de meilleure qualité est devenue nulle exactement au moment où vous avez pu disposer de latrines extérieures ? Est-ce que la valeur qu'a pour vous l'éducation est devenue nulle lorsque vous avez terminé l'école primaire ? La seule réponse honnête est « non ».

La règle d'or

Un principe répandu (pour ne pas dire universel) du « réalisme moral » pourrait ressembler à la « règle d'or »[17] (« fais à autrui ce que tu voudrais qu'il fasse pour toi ») ou à l'impératif catégorique d'Emmanuel Kant (« agis uniquement d'après la maxime qui fait que tu puisses vouloir en même temps qu'elle devienne une loi universelle », KANT, 1998 [1785]). Selon la règle d'or/l'impératif moral de Kant, et tout simplement le bon sens, adopter pour l'évaluation générale du bien-être des autres un standard que vous n'accepteriez jamais pour vous-même est moralement inacceptable.

Toute tentative pour « trancher » la question en prétendant que la fonction-objectif est une « combinaison » d'objectifs fétichistes et non fétichistes implique que les objectifs généraux sont non fétichistes et qu'il s'agit juste une question de pondération, mais les gains importants au-dessus du seuil sont pertinents. Le remplacement des OMD « fétichistes » par les ODD, d'une portée plus large, devrait avoir mis un terme à la pertinence et à la légitimité de l'approche fétichiste comme expression légitime du développement (PRITCHETT, 2015).

Allégation nécessaire, mais fausse numéro 3 : les RCT peuvent apporter des preuves fiables permettant d'améliorer les programmes ciblés destinés aux objectifs de développement « fétichistes » (agrégés ou spécifiques)

Un autre argument pour faire valoir que les RCT constituent le « meilleur investissement » est d'affirmer qu'une évaluation de l'impact de programmes/projets s'appuyant sur des RCT est susceptible de fournir des preuves rigoureuses, fiables et exploitables permettant de concevoir des programmes plus efficaces. Comme bien d'autres experts, dont certains auteurs du présent ouvrage, j'ai expliqué : (a) que ces allégations n'ont jamais été étayées par des preuves solides, mais relèvent seulement d'une conviction ; (b) que les allégations selon lesquelles les RCT pourraient « trancher les débats » sur les impacts liés à l'hétérogénéité des études non expérimentales étaient *ex ante* non seulement empiriquement improbables, mais également impossibles sur le plan logique (PRITCHETT et SANDEFUR, 2013a) ; (c) que l'examen des études empiriques n'a pas réussi à mettre en évidence une cohérence suffisante pour conclure à leur fiabilité (VIVALT, 2020), même sur des sujets spécifiques comme l'amélioration des apprentissages au niveau de l'enseignement de base (EVANS et POPOVA, 2016)

17. PARFIT (2011) argue que trois approches communes des questions morales – la morale déontologique de Kant, le conséquentialisme et le contractualisme – convergent finalement vers les mêmes réponses et que celles-ci sont les réponses « justes ».

ou la vermifugation (notamment les « *Worm Wars* »), et que la variabilité entre les études « rigoureuses » est suffisante pour que, au moins dans certains cas, le fait de s'appuyer sur des preuves « rigoureuses » ne réduise pas l'erreur de prévision sur l'impact d'un programme dans un contexte donné par rapport à des méthodes simples (Pritchett et Sandefur, 2015), ce qui est exactement le résultat auquel tout le monde s'attendait, sauf les *randomistas* (Pritchett, 2018c) ; (d) que la « validité conceptuelle » (la robustesse des résultats en fonction de « l'espace de conception », soit l'ensemble des dimensions qui peuvent varier pour la mise en œuvre) des RCT est faible (Nadel et Pritchett, 2016 ; Kerwin et Thornton, 2018 ; Kaffenberger, 2018) ; enfin, (e) que l'on ne peut pas utiliser des résultats « prouvés » par un organisme exécutant pour extrapoler l'impact obtenu lors de la mise en œuvre par une autre organisation, en particulier lorsque l'on passe de la « démonstration du concept » par une ONG à l'application à grande échelle au niveau d'un gouvernement (Bold *et al.*, 2018 ; Vivalt, 2020).

L'« étude sur les animaux d'élevage » (Banerjee *et al.*, 2015a) à laquelle le professeur Blattman a fait référence est parfois considérée comme la « preuve » que l'on peut créer un « étalon-or » susceptible de guider l'établissement de programmes anti-pauvreté efficaces. Il convient à cet égard de noter sept points intéressants. Premièrement, le taux de rentabilité interne (TRI) est de 7 %, ce qui n'est pas particulièrement remarquable ; il ne dépasserait pas le taux de rendement de 10 % traditionnellement appliqué pour l'analyse coûts-bénéfices de projets de la Banque mondiale. Deuxièmement, j'ai été généreux dans les calculs présentés précédemment en n'y incluant pas les résultats de l'un des six pays, à savoir le Honduras, dans lequel les animaux en question (les fameux poulets !) sont morts, le programme ayant de ce fait eu des effets négatifs assez importants pour les ménages. La moyenne indiquée ne prend donc pas en compte l'ensemble des expériences. Troisièmement, il n'apparaît pas clairement que ce programme puisse être supérieur à un transfert monétaire, étant donné le montant élevé des coûts engagés pour produire les gains constatés. Quatrièmement, il n'y a (encore) aucune preuve « rigoureuse » que les gains du programme se poursuivront durablement. Les calculs suggérant que ce programme permet d'obtenir une VAN positive reposent sur l'*hypothèse* que les gains de l'année 3 perdureront à long terme, les données fournies ne permettant pas de confirmer cette hypothèse. Si l'on considère la baisse observée dans la consommation annuelle de biens durables mesurée entre l'année 2 et l'année 3, et que l'on extrapole les flux de revenus futurs sur la base de ce déclin, on obtient une VAN *négative* pour quatre des six pays. Cinquièmement, Bauchet *et al.* (2015) ont réalisé une évaluation d'impact pour un programme très similaire dans le sud de l'Inde, qui n'a révélé aucun effet sur les revenus ou les biens. Ils expliquent que, dans une période où l'économie locale enregistrait une croissance solide, l'option des animaux d'élevage n'était pas attractive. Nous savons donc déjà avec certitude que ces résultats manquent de validité externe, du moins dans certaines conditions extérieures. Sixièmement, on peut douter de la « validité conceptuelle », dans le sens où ce programme « multidimensionnel » était

complexe et comportait de nombreux éléments, en partie parce que sa conception était le fruit d'un long processus plus informel « d'essais et d'erreurs » et « d'apprentissage par l'expérience » (PRITCHETT *et al.*, 2012), mené par le Building Resources Across Communities (BRAC) et, que par conséquent, même des variations mineures dans la conception ou la fidélité au protocole de mise en œuvre risquaient de ne pas aboutir à des résultats positifs. Septièmement, comme l'étude a été réalisée dans plusieurs pays par une même organisation qui était responsable de son déploiement sur l'ensemble des sites, les autres organisations pourraient être amenées à douter de la robustesse des résultats.

La pertinence de ces remarques au regard de ce chapitre est que, si l'on veut prétendre que le « meilleur investissement » consiste à réaliser une étude dans un domaine qui produit des bénéfices extrêmement limités (comme la conception de programmes ciblés spécifiques à un secteur) par rapport à d'autres études apportant des bénéfices très importants (comme la promotion du développement national), les gains compensatoires dans la probabilité d'obtenir des résultats fiables et exploitables doivent être considérables. Si les recherches sur le développement national ont une chance sur mille de produire des résultats exploitables et les RCT 100 % de chances, alors c'est un argument de poids en faveur des RCT (ou de certaines). Il n'existe toutefois aucun argument convaincant ou preuve irréfutable que la probabilité qu'un travail d'une ampleur donnée engagé dans des RCT donne des résultats fiables et exploitables soit *supérieure*, et encore moins bien supérieure.

Allégation nécessaire, mais (probablement) fausse numéro 4 : les connaissances que les RCT peuvent produire constituent une contrainte essentielle pour l'adoption et la mise en œuvre de meilleurs programmes ciblés

Si on veut qu'une politique/un programme/un projet public ait un impact (durable), elle/il doit satisfaire trois conditions : (1) être « correct(e) sur le plan technique » (en cas de mise en œuvre, elle/il doit prendre pour base un ensemble approprié de liens de causalité entre les *inputs*, les réalisations et les résultats) ; (2) être « défendable sur le plan politique » (il faut être en mesure de créer et de maintenir une alliance politique ayant un pouvoir suffisant pour autoriser les actions et ressources nécessaires) ; enfin (3) être « réalisable sur le plan administratif » (on doit être capable, avec la capacité administrative disponible (ou celle qui peut être mobilisée ou créée) de mettre en œuvre le programme avec une fidélité au protocole suffisante pour atteindre les résultats). Les allégations relatives aux améliorations du bien-être humain issues des connaissances acquises par le biais des RCT sont liées aux affirmations selon lesquelles les connaissances sur la conception des programmes que les RCT peuvent générer constituent « la » (ou tout du moins « une ») contrainte majeure par rapport à d'autres pour une action efficace (PRITCHETT, 2018b). Mais il n'est pas évident que la conception d'une politique soit déterminante pour les résultats. CHONG *et al.* (2014) montrent, pour la mesure du résultat d'une politique très spécifique – c'est-à-dire le renvoi à l'expéditeur de courriers envoyés à des adresses inexistantes à l'étranger – (a) que

la politique appliquée de droit est exactement la même dans tous les pays ; (b) que le résultat – à savoir le pourcentage de courriers retournés conformément à la politique de droit – varie dans la plage maximale possible (c'est-à-dire de 0 % à 100 %) ; et donc (c) que ces variations découlent intégralement des conditions de mise en œuvre, et non de la politique elle-même.

L'« espace de conception » d'un projet/programme destiné à atteindre un objectif (par exemple, l'émancipation des femmes, la réduction des variations des revenus des agriculteurs, l'augmentation de l'épargne, la diminution de la morbidité liée aux maladies hydriques, etc.) est susceptible d'être large et complexe (voire inconnu), dans la mesure où il existe de nombreux choix (par exemple, décider qui est responsable de quelles actions, quelle doit être la fréquence des visites, quel est le contenu des messages informatifs transmis, quelle est la portée d'un prêt, etc.) et plusieurs possibilités pour chaque choix, certains éléments déterminants pour la réussite de l'expérience pouvant même ne pas être connus au stade de la conception. La réalisation d'une RCT permet d'établir une estimation de « l'impact », représenté par un point (ou un ensemble de points, un pour chaque groupe de traitement) sur la « surface de réponse » des résultats pour une conception particulière. La partie précédente montrait à quel point la conclusion tirée d'un point ou d'un ensemble de points est utile (ou non) lorsque la surface de réponse peut varier selon les contextes ou être très inégale (non robuste) au niveau de la conception de la RCT. Mais on peut également craindre que le fait de connaître la surface de réponse relative à la conception d'un projet/programme, qui est impossible sur le plan administratif ou politique, ait une valeur limitée, voire nulle[18] (Gass et Pritchett, 2017 ; Pritchett, 2018a).

Des connaissances disciplinaires pourraient être apportées du seul fait de savoir que des projets/programmes auront un impact X, Y ou Z s'ils étaient adoptés, même si la probabilité de leur adoption demeure nulle. Hormis ces connaissances, ces programmes ne peuvent prétendre apporter des bénéfices en termes de bien-être humain. En s'appuyant sur la publication de Filmer et Pritchett (1999), Pritchett (2010b) soutient que le plaidoyer sur l'utilité des RCT pour « l'élaboration de politiques » repose pour beaucoup sur un modèle politique « *normative as positive* », même dans des domaines où ce modèle s'est révélé manifestement erroné. Il n'est pas nécessaire d'adhérer totalement à la théorie des choix publics pour accepter que l'on ne puisse pas prendre au sérieux, en tant que modèle positif, l'idée que les acteurs publics (hommes politiques, décideurs, hauts technocrates) optimisent une fonction de bien-être social en étant contraints uniquement par leurs connaissances de « ce qui fonctionne ».

La même logique est valable pour la capacité des organisations censées mettre en œuvre des programmes. Savoir qu'un programme aurait un impact X s'il pouvait être mis en œuvre fidèlement ne veut pas dire que les organisations présentes dans le pays, qu'elles soient publiques ou privées, en ont la capacité.

18. Il s'agit juste de la condition évidente de Kuhn-Tucker découlant de l'optimisation (potentiellement) soumise à différentes contraintes ; le « lagrangien » ou « prix fictif » sous contraintes faibles est nul.

Bon nombre des RCT existantes n'ont pas pu démontrer le lien de causalité entre la conception de l'intervention, les réalisations de l'organisation responsable de leur mise en œuvre et les résultats. Au lieu de cela, il est apparu que, même dans le cadre restreint d'une expérience, le « traitement » (qu'il s'agisse d'une rémunération au rendement, d'une information aux citoyens ou des injonctions descendantes) n'a pas contribué à modifier le comportement des agents exécutants destiné à produire des « résultats », comme le montrent par exemple Banerjee *et al.* (2008) (aides-soignantes et sages-femmes au Rajasthan), Banerjee *et al.* (2010) (directeurs d'école publique, enseignants en Uttar Pradesh), Banerjee *et al.* (2012) (police au Rajasthan). On peut également citer des exemples d'expérimentations qui ont produit des résultats lors de leur mise en œuvre par une ONG, et qui n'ont pas « fonctionné » lorsqu'elles ont été appliquées à grande échelle par les gouvernements, par exemple, Duflo *et al.* (2012) sur l'utilisation de caméras dans les salles de classe d'écoles rattachées à des ONG, et Dhaliwal et Hanna (2013) sur la biométrie dans les centres médicaux publics dans le Karnataka. L'ampleur de l'écart observé entre les pays au niveau des programmes anti-pauvreté ou sectoriels s'explique par des différences notables dans les capacités de mise en œuvre des différents pays, plutôt que par la conception des politiques. Le tabl. 2 montre les variations importantes dans les indicateurs nationaux de la capacité de l'État et leurs liens avec les résultats mesurés sur le bien-être humain entre les pays.

Selon d'assez bons arguments, les connaissances « techniques » ou « codifiables » que les RCT sont les mieux à même de produire constituent au mieux une contrainte mineure dans l'adoption et la mise en œuvre efficace de programmes ciblés (Pritchett, 2018a ; 2018b) par rapport aux contraintes politiques qui influencent la « volonté de faire » et la capacité à « pouvoir faire », qui ne sont quant à elles aucunement impactées par les résultats des RCT. En revanche, on peut aisément affirmer que l'exploitation des connaissances existantes dans un pays donné est un facteur endogène de la politique et de la capacité, plutôt qu'un facteur exogène, la partie « codifiable » des connaissances étant un bien public qui, n'étant ni rival ni exclusif[19], devrait se diffuser rapidement et facilement.

Conclusion

Une évaluation d'impact réalisée par le biais d'une RCT ne paraît pas du tout être le bon outil pour les pays et les gouvernements, ou pour les agences qui souhaitent promouvoir le développement. Au lieu de cela, il semble surtout que

19. En économie, on peut classer les biens en fonction de deux critères : la notion de « rivalité » d'usage entre consommateurs – la consommation d'un bien par un consommateur limitant la capacité des autres à consommer le même bien – et la notion d'« exclusion », qui renvoie à la capacité de s'approprier un bien en en payant le prix.

cette méthode serve à guider cette petite partie du processus de développement que représentent les activités « caritatives » ou « philanthropiques » et qui (a) distribuent des sommes d'argent relativement faibles ; (b) ne passent pas ou ne peuvent pas passer par les gouvernements nationaux (ou régionaux ou locaux) ; (c) ont des évaluations relativement « fétichistes » (peut-être en partie parce qu'elles rationnent des ressources infimes) ; enfin (d) se soucient surtout de pouvoir attribuer le gain de bien-être directement à leur intervention spécifique (plutôt qu'à des effets indirects). Les œuvres caritatives sont une bonne chose, et il paraît utile qu'elles puissent être mieux orientées par les données résultant des RCT. C'est probablement ce travail caritatif que Bill Gates et Chris Blattman avaient en tête lorsqu'ils en sont venus à évoquer les poulets et leur impact.

Mais il n'est pas du tout crédible d'assimiler ce minuscule segment du monde au processus de développement plus large. Si l'actuelle Corée du Sud ne ressemble pas à celle du début des années 1960, c'est parce que son gouvernement ne s'est pas contenté de promouvoir la possession et l'élevage de volailles. Le monde d'aujourd'hui est radicalement meilleur sur la quasi-totalité des indicateurs objectifs du bien-être humain grâce au développement national général (qui inclut la croissance économique) et à l'amélioration de la performance sectorielle, qui sont justement censés constituer l'objet du développement (Pritchett, 2017). Imaginer que les outils promus par les ONG internationales pour identifier les actions humanitaires efficaces au profit des plus pauvres – en particulier lorsque les actions en question se limitent à celles dont les effets sont directement imputables à l'intervention de ces mêmes ONG – sont aussi le « meilleur investissement » pour réduire la pauvreté (sans parler du meilleur investissement pour le développement), ce n'est pas promouvoir rationnellement une méthode, mais de la folie pure.

Chapitre 3

Le pouvoir subversif des expérimentations aléatoires

Jonathan MORDUCH

Introduction

Il existe deux façons distinctes d'utiliser les évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) en économie du développement. La première se sert des RCT pour mesurer l'impact. Avec la seconde, les RCT permettent d'explorer la nature des contrats économiques, des comportements et des institutions. Les critiques ont souvent tendance à assimiler ces deux types de RCT, alors que chacun d'eux aborde des problématiques très différentes et vise des objectifs distincts. Pour comprendre le pouvoir des RCT, et s'y retrouver dans les débats autour des évaluations randomisées, il faut d'abord bien distinguer les deux types.

Si les détracteurs sont particulièrement mal à l'aise avec le fait d'élever les RCT au rang d'outil d'évaluation privilégié, on peut accepter totalement ou partiellement leurs critiques et reconnaître malgré tout l'importance des RCT (et vouloir encourager la réalisation d'un plus grand nombre de RCT) pour l'expérimentation. Les *randomistas* doivent-ils faire la loi (Ravallion, chap. 1, ce volume) ? Non. Les RCT constituent-elles un étalon-or (BÉDÉCARRATS *et al.*, 2019a) ? Non. Dans la pratique toutefois, les RCT ont été (et continueront à être) des outils exploratoires extrêmement utiles.

Le premier usage des RCT (et la cible des critiques les plus ferventes) est la réalisation d'évaluations d'impact au moyen de méthodes randomisées. La critique vise moins souvent la nature même de ces RTC que le fait de les mettre sur un piédestal et de leur conférer un statut particulier qui leur accorde une plus grande crédibilité que d'autres méthodes d'évaluation. Ces RCT sont

axées sur l'évaluation des programmes et des politiques mis en œuvre par les gouvernements et les ONG, et leurs partisans caressent l'espoir que l'obtention de mesures d'impact plus crédibles *via* la randomisation permettra d'améliorer les investissements et les actions (GLENNERSTER et TAKAVARASHA, 2013 ; KREMER, 2003 ; BANERJEE et DUFLO, 2009). Les questions portent généralement sur « ce qui fonctionne ». Des études antérieures incluent des RTC sur des programmes gouvernementaux comme le programme de transferts monétaires conditionnels *Progresa* au Mexique (LEVY, 2006) et le *Job Training Partnership Act* (loi sur le partenariat dans la formation professionnelle) aux États-Unis (LALONDE, 1986), et la vague la plus récente compte des évaluations de programmes d'ONG comme des RCT sur le microcrédit, qui sont regroupées dans la publication de BANERJEE *et al.* (2015c). Dans la majorité des cas, les chercheurs conçoivent les évaluations, mais pas les actions. L'enjeu principal de cet ouvrage consiste à savoir si et de quelle façon les preuves empiriques par randomisation devraient être prises en compte.

Le second type de RCT est d'une nature différente. Il s'inscrit dans une approche expérimentale qui connaît un succès croissant auprès des économistes du développement et qui met en avant les RCT comme une innovation méthodologique essentielle. Alors que certaines expérimentations économiques exploitent des scénarios hypothétiques élaborés en laboratoire (par exemple DAVIS et HOLT, 1993), cet axe des RCT s'appuie sur des expériences en conditions réelles. Les études reposent sur des manipulations contrôlées expérimentalement de structures de prix, contrats, méthodes d'enseignement, protocoles de soins, processus bureaucratiques et autres. Dans ce cas, les chercheurs participent activement à la conception des programmes et des politiques réels, généralement en coopération avec un organisme gouvernemental, une entreprise ou une ONG. Les questions posées ont une visée exploratoire, sont basées sur la théorie et motivées par la volonté de cerner les possibilités et les contraintes économiques. Les contextes de réalisation sont souvent des pilotes à petite échelle ou des essais sur une durée limitée. Les questions portent moins souvent sur « ce qui fonctionne » que sur « comment et pourquoi cela fonctionne » ou sur « que pourrait-on faire ? ». Alors que les RCT d'évaluation sont critiquées parce qu'elles donnent peu d'informations sur le « pourquoi » (pourquoi tel impact est-il faible ou important ? pourquoi apparaît-il pour certaines personnes, mais pas pour d'autres ?), cet autre type d'études est axé sur l'explication. Elles conduisent en fin de compte à se demander si la marche du monde est bien celle préconisée par la théorie économique. Le pouvoir de ces RCT réside dans la façon dont elles viennent perturber les conceptions courantes en manipulant les environnements économiques, révélant ainsi au grand jour ce qui, sinon, resterait caché ou inexploré.

La frontière entre les deux types de RCT peut être bien floue, troublée par ce qu'en disent à la fois les partisans et les opposants de ces expérimentations, et l'objectif de ce chapitre est d'expliquer les deux méthodes et d'illustrer l'approche expérimentale en économie du développement. La vision que je propose n'est pas forcément celle qui serait défendue par un *randomista* pur et

dur, mais elle correspond à la façon dont les RCT sont souvent utilisées dans la pratique[1].

La première partie de ce chapitre décrit la montée en puissance de l'approche expérimentale associée aux RCT. La deuxième partie présente trois exemples de RCT qui explorent des questions relatives aux prix, aux contrats et à l'utilisation de services financiers dans des communautés pauvres. La troisième partie traite de l'intérêt des RCT pour les actions visant à réduire la pauvreté et à fournir des biens privés, et étudie l'argument selon lequel les RCT détournent l'attention de l'étude des forces systémiques qui façonnent les économies.

Étendre les connaissances en créant des variations

On peut craindre que la primauté accordée aux RCT à des fins d'évaluation – avec des méthodes associées comme les expériences naturelles et les régressions sur discontinuité – et leur montée en puissance dévalorisent indûment et inutilement d'autres méthodes visant à évaluer les programmes qui fonctionnent (notamment la régression linéaire, les variables instrumentales classiques, l'ethnographie et l'évaluation qualitative, et l'apprentissage automatique s'appuyant sur le *big data*). Plus préoccupant encore : donner la priorité à ce type de RCT pourrait limiter l'attention seulement aux actions économiques qui se prêtent bien aux évaluations randomisées. Et la crainte ultime est que le statut spécial accordé aux RCT pour déterminer « ce qui fonctionne » pourrait entraîner une perte de connaissances, notamment de celles que l'on pourrait retirer d'approches diverses (Ruhm, 2019).

Les détracteurs des RCT s'inquiètent aussi du fait que leurs défenseurs exagèrent la précision et la facilité de généralisation des évaluations randomisées (Deaton et Cartwright, 2018). Ils ont peur que le type d'informations évaluatives générées par les RCT ait souvent une valeur limitée sur le plan politique et pratique (Pritchett, 2014c ; Drèze, 2018b), tout en étant exposé aux interprétations erronées faute de contextualisation (Morvant-Roux *et al.*, 2014). Comme d'autres méthodes d'évaluation, les RCT ont des difficultés à fournir des réponses précises, notamment quand il est nécessaire (et c'est souvent le cas) de transposer les données d'une étude réalisée à un endroit dans l'environnement politique d'un autre (Cartwright et Hardie, 2012 ; Pritchett et Sandefur, 2015 ; Bisbee *et al.*, 2017)[2].

Ce qui est peut-être le plus inquiétant, selon l'argument des critiques, c'est que les actions susceptibles de se prêter le mieux à une évaluation par des RCT

1. Voir Ogden (2017), qui présente la vision d'universitaires et de praticiens participant à des RCT, notamment sur le sujet des diverses théories du changement.

2. Comme le fait remarquer Imbens (2018), les chercheurs qui utilisent les RCT sont toutefois conscients des limites, et répondent par des approches élargies (voir par exemple Bates et Glennerster, 2017).

sont trop limitées, de trop faible ampleur et trop spécifiques. En économie, les RCT sont mieux adaptées aux études concernant les biens privés que les biens publics. En outre, les RCT sont souvent centrées sur des impacts marginaux et les impacts sur des sous-populations marginales (Wydick, 2016). Elles peuvent ainsi être utilisées pour mesurer un impact à court terme, par exemple lorsque l'on introduit le microcrédit dans une nouvelle région, mais pas pour évaluer comment ses clients initiaux s'en sortent depuis qu'il a été mis en place (Cull et Morduch, 2018).

Sur un plan plus large, en se focalisant sur les petits pas vers l'amélioration de la mise en œuvre de concepts existants, les données d'impact issues des RCT tendent à fournir seulement des informations indirectes sur les structures plus importantes qui perpétuent la pauvreté et les inégalités. Dans cette perspective, donner la primauté absolue aux RCT en matière d'évaluation contribuerait à limiter les preuves admissibles de « ce qui fonctionne », et à restreindre finalement la compréhension de phénomènes économiques et sociaux complexes (Bédécarrats *et al.*, 2019a).

A contrario, l'usage des RCT à des fins exploratoires (le second type de RCT précédemment cité) élargit clairement les connaissances et la plupart des RCT qui ont fait l'objet de publications par des économistes du développement vont dans ce sens. L'approche expérimentale permet de remédier au fait que les variables clés n'ont pas naturellement tendance à évoluer beaucoup, les expériences servant ainsi à créer des changements pertinents. Il est possible que les prix ou les contrats n'enregistrent pas de variation majeure à un moment donné ou dans un échantillon donné. Il est possible que les gouvernements, les cliniques ou les écoles agissent de manière uniforme, dans une certaine mesure. Résultat : même si les chercheurs peuvent étudier des prédictions théoriques, ils ont peu d'espoir d'arriver à les tester de manière empirique. Sans passer par l'expérimentation, il y a bien peu d'éléments à observer et donc bien peu de choses à analyser.

Mais ces RCT exploratoires ont, elles aussi, des limites. Il est en effet très tentant de tirer des conclusions politiques exagérément solides à partir d'essais et d'expériences pilotes, plutôt que de les considérer simplement pour ce qu'elles sont, c'est-à-dire informatives et provocantes, mais contingentes. Dans le même temps pourtant, en critiquant ces RCT parce qu'elles sont des expériences pilotes ou des essais, on risque de négliger le fait qu'elles peuvent s'additionner pour offrir une compréhension plus fine et plus étendue des contraintes et des possibilités. Même si Angus Deaton et Nancy Cartwright sont contre le fait de conférer un statut particulier aux preuves issues des RCT, ils font remarquer que :

> « Les RCT sont souvent un moyen commode pour introduire une variation contrôlée expérimentalement : si vous voulez voir ce qui se passe, alors donnez un coup de pied dans la fourmilière et constatez le résultat, tirez la queue du lion[3] ! » (Deaton et Cartwright, 2018 : 17).

3. N.D.E. : l'expression « tirer la queue du lion » renvoie à l'idée de provoquer, énerver, caresser à rebrousse-poil.

Sur le plan des connaissances économiques, se servir des RCT pour « tirer la queue du lion » a permis aux chercheurs de mieux comprendre les théories économiques, et les a poussés à remettre en question des hypothèses qui étaient considérées comme définitivement établies.

Prenons le cas de l'assurance récolte : il s'agit d'un produit à fort potentiel étant donné les risques liés à l'agriculture pluviale. Pourtant, dans la pratique, cette assurance (tout comme sa variante plus récente, l'assurance indicielle basée sur la pluviométrie) a été extrêmement difficile à vendre aux agriculteurs. La publication de CASABURI et WILLIS (2018), par exemple, montre que seulement 5 % des planteurs de canne à sucre kényans dans leur échantillon ont souscrit une assurance précipitations. Ce résultat confirme le sentiment que les clients potentiels se méfient de ces produits, ne les comprennent pas forcément ou ne leur font pas confiance, ils se contentent de systèmes informels et/ou trouvent que ces produits sont mal conçus ou trop coûteux. Casaburi et Willis ont néanmoins recours à une RCT pour tester le moment choisi pour vendre cette assurance. Ils se demandent si le problème se situe vraiment au niveau des prix ou de la compréhension des clients, ou si le faible taux de souscription ne pourrait pas venir plutôt du fait que les assureurs demandent que la prime soit payée en une fois et avant la saison de plantation, à une période où les producteurs investissent l'essentiel de leur argent dans les plants. En randomisant la période de paiement et en reportant celui-ci au moment des récoltes (lorsque les producteurs ont des liquidités) pour un échantillon des clients, ils obtiennent une augmentation du taux de souscription, celui-ci passant à 72 %. En revanche, une réduction de 30 % du coût de l'assurance (sans reporter la période de paiement) a augmenté la demande d'un point de pourcentage seulement. La RCT a permis de conserver les autres conditions sans les modifier et, même si son résultat n'est pas révolutionnaire, elle a contribué à mieux percevoir le problème. Il est moins important de savoir si le paramètre est transposable à l'identique ou non que d'avoir pu identifier, grâce à l'étude, que la période de paiement et les liquidités sont des contraintes pour la demande d'assurance et qu'elles doivent être prises en compte sérieusement dans d'autres contextes (outre le fait d'avoir pu apporter une réponse pratique au problème)[4].

L'expérience menée par Casaburi et Willis au Kenya a guidé les travaux de BELISSA *et al.* (2019) en Éthiopie. Ils étudient également le rôle des liquidités dans la souscription de l'assurance, en se demandant eux aussi si la demande est plus élevée quand les agriculteurs peuvent payer après la récolte, lorsqu'ils disposent de plus d'argent liquide. Ils analysent également l'impact de la promotion de l'assurance par le biais des *Iddirs*, des dispositifs de partage des risques informels utilisés par les fermiers en Éthiopie. La RCT réalisée par BELISSA *et al.* (2019) porte sur 8 579 personnes et 144 *Iddirs*, répartis dans six groupes. Le premier est

4. De manière analogue, Jonathan Bauchet et moi-même avons étudié la demande d'assurances-vie vendues à des femmes pauvres au Mexique. En mettant en œuvre une expérience naturelle, nous avons établi que la demande augmentait de plus de 59 % lorsque les clientes étaient autorisées à payer sous la forme de petits versements hebdomadaires plutôt qu'en un versement unique (BAUCHET et MORDUCH, 2019).

un groupe de contrôle auquel on propose un contrat standard d'assurance indicielle basée sur la pluviométrie, dont la prime doit être payée avant la prise d'effet du contrat. Le deuxième groupe est similaire, excepté le fait que c'est un responsable local qui fait la promotion du produit. Le troisième groupe est également comme le premier, mais des paiements différés sont autorisés. Le quatrième est similaire au troisième, mais on demande au souscripteur de signer un engagement contraignant à verser la prime après la récolte. Dans le cinquième groupe, l'*Iddir* fait la promotion de l'assurance (avec l'option de paiement différé), et le sixième groupe réunit tous ces critères : option de paiement différé, obligation de signer un engagement contraignant et promotion de l'assurance par le biais de l'*Iddir*.

Même si le phénomène est moins spectaculaire que dans l'étude de Casaburi et Willis, le fait de différer la période de paiement s'avère être un facteur important pour les fermiers éthiopiens, puisqu'il fait passer le taux de souscription de 8 % à 24 %. En combinant le paiement différé et la promotion par les *Iddirs*, on renforce l'impact, le taux de souscription passant à 43 %. La promotion de l'assurance par les *Iddirs* n'apporte pas seulement de la crédibilité au produit, mais favorise aussi l'achat collectif de l'assurance dans un contexte explicite d'assurance informelle. Mais l'étude révèle aussi que 15 % environ des agriculteurs qui avaient accepté de payer après la récolte ont en fait manqué à leur engagement, ce chiffre étant suffisamment élevé pour mettre en péril la viabilité économique du produit d'assurance.

Ces deux études sur l'assurance illustrent la distinction fondamentale entre les RCT utilisées à des fins exploratoires (c'est-à-dire des expériences conçues par des chercheurs qui font émerger des pistes d'investigation des systèmes mis en place) et celles servant à évaluer l'impact de programmes établis. Aucune des deux études décrites ici ne mesure l'impact de l'assurance sur les agriculteurs et aucune n'intègre un groupe de contrôle qui ne serait bénéficiaire d'aucune action. L'objectif principal n'est pas d'évaluer si l'assurance « fonctionne ». Au lieu de cela, le groupe de contrôle a, dans les deux études, la possibilité d'acheter un produit d'assurance standard. Les deux études étudient ensuite ce qui se produit lorsque les produits sont repensés de façon systématique pour évaluer le comportement des fermiers et la viabilité des produits. Aucun des résultats spécifiques des deux expériences ne peut être extrapolé dans d'autres contextes. Mais ce qui peut l'être, ce sont la nature des innovations apportées (période de paiement différé, marketing par le biais de groupes locaux) et les préoccupations générales liées au produit (illiquidité, risque de non-paiement après la récolte).

En matière d'évaluation d'impact, les RCT sont souvent mises en avant parce qu'elles réduisent le biais de sélection lié à un accès non aléatoire aux programmes. Les deux cas sur l'assurance montrent cependant que le biais de sélection n'est qu'un grand défi parmi d'autres dans l'économie du développement empirique. L'un des principaux problèmes ici est l'insuffisance de variation pertinente dans les contrats d'assurance (notamment l'absence de contrats proposant des options de paiement après la récolte), problème que révèle l'expérimentation par l'intermédiaire de la RCT. Ni l'une ni l'autre des études ne devait nécessairement être une RCT, mais elles étaient tenues toutes les deux d'intégrer une expérimentation et une redéfinition du produit. Les deux étaient

obligées de « tirer la queue du lion » et bousculer l'ordre établi. Le choix des deux groupes de chercheurs en faveur des RCT s'explique par la commodité à associer expérimentation et randomisation dans une méthode de type exploratoire.

Alors que Ravallion (chap. 1, ce volume) fait remonter le recours aux RCT en économie à des expériences menées dans les années 1950 et 1960 (Gueron, 2017), la montée significative des RCT dans le domaine de l'économie du développement a démarré dans les années 1990, après une période de maturation méthodologique qui, entre autres résultats, a conduit à s'intéresser aux expériences naturelles (Angrist et Krueger, 1999). Sur le plan conceptuel, l'évolution des expériences naturelles vers les RCT, initiée au Kenya par Michael Kremer de l'université de Harvard, s'est faite en douceur. Elle s'est intensifiée plus tard avec la création du Abdul Latif Jameel Poverty Action Lab (J-PAL) au MIT (Massachusetts Institute of Technology) (Kremer, 2003 ; Banerjee et Duflo, 2009[5]). Michael Kremer et ses collègues ont participé à la conception des actions, ce qui n'était pas le cas pour les RCT évaluatives antérieures mises en œuvre pour tester des interventions développées au niveau gouvernemental. Kremer (2003) résume une série d'expérimentations initiales menées au Kenya afin d'améliorer les résultats en matière de scolarisation, qui incluaient notamment la fourniture de petits-déjeuners gratuits, d'uniformes scolaires et de manuels, une opération de déparasitage des enfants et l'embauche d'enseignants supplémentaires. Plusieurs de ces actions ont permis d'augmenter sensiblement le taux de fréquentation des écoles à un coût relativement faible.

Les exemples présentés montrent bien l'origine de la confusion entre les types de RCT. Kremer (2003) décrit les RCT comme des méthodes d'évaluation de « ce qui fonctionne » (dans le sens mentionné plus haut). Pourtant, sans rien ôter à leur valeur, elles sont par nature exploratoires. Il s'agit en effet en grande partie de programmes pilotes, et non d'actions publiques déployées à grande échelle. Elles fournissent des informations précieuses sur les possibilités et les contraintes, ouvrent des perspectives importantes et constituent, non pas un point final, mais un pas vers de nouvelles étapes.

L'omniprésence de la sous-optimalité et le potentiel d'innovation

Deaton et Cartwright (2018) veillent à bien distinguer les RCT de type « évaluation de ce qui fonctionne » et les RCT permettant d'explorer « le pourquoi et le comment »[6]. Dans cette perspective, ils prennent en compte les cas

5. Voir Ogden (2017) pour une description de la méthode et des motivations de ces trois auteurs.

6. Deaton et Cartwright (2018) présentent un numéro spécial de la revue *Social Science and Medicine* sur le thème « Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue », édité par I. Kawachi, S.V. Subramanian et R. Mowat, qui traite 19 réponses émanant de statisticiens et de chercheurs en sciences sociales de premier plan.

« où les RCT parlent d'elles-mêmes », et les situations où « aucune extrapolation ou généralisation n'est requise » :

> « Pour certaines choses que nous voulons savoir, une RCT suffit à elle seule. Elle peut fournir un contre-exemple à une proposition théorique d'ordre général, soit à la proposition elle-même (test d'infirmation simple), soit à une conséquence de celle-ci (test d'infirmation complexe). Une RCT peut également confirmer une prédiction liée à une théorie, et, même si cela ne confirme pas la théorie elle-même, c'est une preuve qui joue en sa faveur, en particulier si la prédiction semble par nature peu probable au départ » (DEATON et CARTWRIGHT, 2018 : 13).

La plupart des RCT exploratoires visent rarement à infirmer une théorie, selon l'acception de Deaton et Cartwright. Les deux cas exposés concernant les assurances, par exemple, sont axés sur des concepts bien connus (l'illiquidité et le manque de confiance), dont l'importance n'a rien de surprenant (dans le sens où il est fort probable qu'ils figurent parmi les défis à relever dans la vente d'assurances). Ce qui est finalement plutôt en jeu, c'est de savoir quelle crédibilité accorder à l'optimisation sous contraintes. Un principe fondamental de l'économie néoclassique est l'idée que les marchés produisent des institutions, des biens et services et des prix optimaux. En théorie, la fonction « disciplinante » du marché doit éliminer les formes sous-optimales. Ce principe continue même à s'appliquer en situation d'optimalité de deuxième et de troisième niveau avec des contraintes comme une information asymétrique et une exécution imparfaite des contrats (STIGLITZ et WEISS, 1981). La théorie économique moderne énonce en substance que ce que nous voyons n'est pas forcément parfait, mais que c'est ce que nous pouvons avoir de mieux[7]. En d'autres termes, les processus et produits d'assurance existants devraient déjà intégrer des moyens pour faire face, dans la mesure du possible, aux problèmes de liquidité et de confiance.

Mais est-ce généralement vrai ? L'expérience menée par Muhammad Yunus sur les contrats de crédit dans les années 1970, qui a conduit au développement du microcrédit, illustre un cas où l'ajustement et la redéfinition des contrats ont apporté une amélioration réelle de ce qui était fourni par le marché. Les nouveaux contrats ont conduit à une baisse considérable des taux d'impayés sur les prêts, et permis d'atteindre une certaine rentabilité, même lorsque les prêteurs appliquaient des taux d'intérêt relativement faibles (ARMENDÀRIZ et MORDUCH, 2010). Ce que les économistes pensaient être un résultat optimal sous contraintes s'est avéré ne pas en être un. Et même les ajustements introduits par Muhammad Yunus n'ont pas été le dernier mot dans les innovations introduites en microfinance (par exemple : RAI et SJÖSTRM, 2004 ; FIELD *et al.*, 2013).

Les RCT exploratoires poursuivent dans cette voie, au travers de l'expérimentation, en contribuant à une cartographie de l'écart entre ce que sont les institutions

7. Un résultat fondamental en économie de l'information est que l'équilibre n'est pas toujours efficace sous contraintes (STIGLITZ, 1986). Le programme de recherche par RCT peut être considéré comme présentant un éventail de circonstances beaucoup plus large de solutions inefficaces.

et les choix existants, d'une part, et ce qu'ils pourraient être, d'autre part. Les RCT permettent de plus en plus de comprendre pourquoi certaines innovations n'ont pas lieu (par exemple, par crainte d'un taux d'impayés relativement élevé, comme rapporté par Belissa *et al.*, 2019), et bien souvent de tester des mesures pratiques pour limiter certains problèmes. Les RCT exploratoires servent rarement à tester un raisonnement théorique spécifique (du type « les individus sont-ils rationnels ? »), mais plutôt à faire la démonstration d'une innovation ou d'une manipulation expérimentale, qui révèle (ou permet de mieux comprendre) la sous-optimalité.

Pourquoi les RCT ?

Au sujet de la vérification des théories, Deaton et Cartwright observent que la généralisation n'est pas toujours la préoccupation majeure. Ils ajoutent que :

> « [la vérification de théories] est un territoire connu, et les RCT n'ont rien d'unique : elles constituent seulement un type de procédures de vérification parmi d'autres » (Deaton et Cartwright, 2018 : 12).

À un niveau de généralité élevé, il est certainement vrai que « les RCT n'ont rien d'unique », effectivement. Il existe bien entendu d'autres méthodes capables de tester la théorie, de démontrer la sous-optimalité, de déranger, de surprendre et d'élargir les cadres économiques. Les méthodologistes s'appuient maintenant sur l'approche expérimentale, en améliorant dans certains cas les RCT de base (Kasy et Sautman, 2019), et en combinant dans d'autres cas randomisation et ethnographie (Duncan *et al.*, 2007). Des méthodes non randomisées peuvent également être utilisées pour analyser des perturbations créées de manière exogène. Mais les RCT, en association avec une démarche expérimentale, se sont avérées particulièrement utiles dans la pratique. Le facteur qui joue en faveur des RCT d'évaluation plaide aussi en partie pour les RCT exploratoires : le biais de sélection. C'est un souci constant et les RCT peuvent contribuer à le maîtriser (tout en créant, il est vrai, d'autres difficultés). L'autre argument consiste à dire que, lorsque vous manipulez *déjà* l'environnement économique dans un mode expérimental, la randomisation semble être un simple prolongement.

Les chercheurs qui utilisent les RCT se demandent pourquoi on souhaiterait employer une autre méthode pour étudier la question qui leur est posée dans le lieu concerné. Pourquoi étudier l'élasticité des prix des moustiquaires imprégnées d'insecticide, par exemple, en utilisant une approche non randomisée, alors qu'il est possible de randomiser les prix ? Dans cet esprit, Deaton et Cartwright (2018) citent une réponse fréquente à leur critique des RCT : « O.K., vous avez souligné certains défauts des RCT, mais d'autres méthodes ont tous ces mêmes défauts, plus des défauts qui leur sont propres » (Deaton et Cartwright, 2018 :

16). Deaton et Cartwright refusent cette riposte, car ils trouvent que le recours aux RCT ne fait que remplacer une série de problèmes par une autre. Comme le souligne par exemple Martin Ravallion (chap. 1, ce volume), on rencontre des problèmes dans les RCT lorsque la non-conformité est sélective ou que l'on est face à une « hétérogénéité essentielle ».

Pourtant, les approches alternatives les plus éminentes de l'inférence causale (notamment l'application de variables instrumentales) se heurtent à des limites bien connues. À défaut d'autre chose, l'histoire de l'économie empirique du développement a montré (1) que le biais de sélection a souvent un effet important, et (2) qu'il est difficile de trouver des variables instrumentales et des expériences naturelles probantes. C'est vrai en économie, et plus encore en économie du développement.

BEAMAN *et al.* (2018b) peuvent nous servir d'illustration. Ils ont conçu une expérience pour mesurer la sélection en matière d'emprunt à travers un échantillon de fermiers au Mali. Leur objectif est d'évaluer les rendements du capital et l'impact du microcrédit dans l'agriculture, en s'intéressant à la probabilité que les fermiers les plus prometteurs soient plus enclins à emprunter que les autres. Sans variation exogène et excluable des prix ou d'autres facteurs externes, l'évaluation au moyen de variables instrumentales n'est pas réalisable. À la place, ils ont donc imaginé une RCT à deux niveaux. Pour avoir une vision de l'étendue du biais de sélection, Beaman *et al.* ont choisi de façon aléatoire 88 villages maliens sur 198, dans lesquels ils ont proposé des prêts par l'intermédiaire d'une institution de microfinance locale. Ils ont ensuite randomisé l'attribution de subventions en capital à un échantillon parmi les 110 villages ne bénéficiant pas des prêts, et à un échantillon de non-emprunteurs dans les 88 villages ayant reçu les prêts. Ils ont ensuite mesuré le rendement du capital à la fois pour les emprunteurs et les non-emprunteurs.

En moyenne, les rendements du capital se sont avérés importants et positifs dans un contexte manifeste de problèmes de liquidité. Les bénéficiaires de subventions en capital (d'un montant d'environ 140 $) issus des 110 villages n'ayant pas reçu de prêts ont augmenté leurs terres cultivées de 8 %, leur utilisation d'engrais de 16 % et la valeur totale des intrants de 15 %. Leur revenu net a en conséquence augmenté de 13 %. Des résultats similaires ont été enregistrés dans les 88 villages ayant profité du microcrédit (ce qui va à l'encontre des résultats négatifs bien connus résumés par BANERJEE *et al.*, 2015c). Toutefois, les agriculteurs qui ont choisi de ne pas emprunter, mais qui avaient accès au microcrédit, ont eu un rendement du capital quasi nul à la marge. Ainsi, comparer les rendements des emprunteurs à ceux des non-emprunteurs – sans tenir compte de l'endogénéité de l'emprunt et de l'hétérogénéité des rendements – reviendrait à surestimer considérablement les rendements nets de l'accès au microcrédit.

Le manque de variables instrumentales probantes tend à être plus important dans les études micro-économiques du développement, car les défaillances des marchés induisent des liens d'interdépendance dans les choix des ménages et

dans les marchés, en particulier dans les contextes informels (voir par exemple Bardhan, 1984 ; Stiglitz, 1986). Il est ainsi plus difficile de trouver des variables que l'on peut exclure sans compromettre la plausibilité du résultat, car, en l'absence de marchés complets, un nombre plus important d'éléments économiques se révèlent être endogènes. Les empiristes qui ont travaillé sur le modèle canonique du ménage agricole (Singh *et al.*, 1986), par exemple, ont exploité une propriété récursive justifiant une analyse de la production qui soit indépendante des variables de consommation. L'inverse ne se vérifie toutefois pas, excluant de fait l'utilisation des variables de production comme instruments pour l'analyse des choix de consommation des ménages producteurs-consommateurs (incluant des agriculteurs et des petits entrepreneurs). Et la fonction de récursivité elle-même dépend d'hypothèses fortes sur la complétude des marchés, dont les marchés d'assurances.

On peut en retirer de nombreuses idées intéressantes, mais beaucoup moins de méthodes convaincantes permettant de mettre à l'épreuve et de tester ces idées, même lorsqu'on est en mesure d'observer des variations naturelles dans l'environnement économique. En outre, l'utilisation de variables instrumentales conduit souvent à des situations dans lesquelles, même si les instruments ne sont pas pleinement probants, les paramètres évalués peuvent toutefois être sensiblement impactés par la méthode des variables instrumentales (VI)[8]. Le cadre *Local Average Treatment Effects* (LATE) aide à comprendre pourquoi : avec des effets de traitement hétérogènes, la méthode des moindres carrés ordinaires (MCO) et les VI évaluent essentiellement des paramètres différents (Imbens et Angrist, 1994 ; Imbens, 2010). Parallèlement aux défis d'identification de variables instrumentales plausibles (Staiger et Stock, 1997), le cadre LATE a remis en question l'apport possible des stratégies exploitant les variables instrumentales. Il est apparu clairement que les différences entre des estimations par MCO et VI ne pouvaient pas être considérées comme résultant uniquement (ou même en grande partie) de la suppression des biais (une interprétation naturelle uniquement dans l'hypothèse d'effets de traitement homogènes). Au lieu de cela, la méthode des variables instrumentales produit des paramètres particuliers qui sont spécifiques à l'interaction entre l'instrument et la variable endogène lorsque les effets de traitement sont hétérogènes (Heckman et Urzua, 2010). Même si les RCT génèrent aussi des paramètres locaux et spécifiques, elles peuvent être interprétées au travers du modèle expérimental. Elles offrent ainsi une méthode d'interprétation qui est souvent plus claire qu'une approche LATE type, basée sur une régression au moyen de variables instrumentales, en particulier si celle-ci ne s'appuie pas sur une expérience naturelle ou si elle utilise des instruments multiples définis de façon continue (Samii, 2016).

8. C'est par exemple le cas pour Pitt et Khandker (1998), avec une évaluation non randomisée bien connue du microcrédit au Bangladesh, qui repose sur l'hypothèse de l'homogénéité des effets de traitement et l'utilisation de formes fonctionnelles particulières pour l'identification, et qui s'est finalement avérée ne pas être rigoureuse, même dans ses conditions propres. Pour un débat critique, voir Roodman et Morduch (2014).

Trois exemples

Afin d'illustrer les RCT à visée exploratoire, je décris ci-après trois exemples d'expériences relatives aux contrats, aux prix, ainsi qu'à l'accès aux marchés et produits financiers.

Les contrats de microcrédit

Le contrat de microcrédit type prend une forme inattendue pour un prêt professionnel. Bien qu'il soit décrit comme un prêt d'investissement pour petite entreprise, ce prêt apparaît plutôt comme un prêt à la consommation, avec des contrats basés sur un remboursement par versements réguliers démarrant peu après son déblocage. Les prêts octroyés par la Grameen Bank font par exemple l'objet de remboursements hebdomadaires qui commencent la semaine suivant la mise à disposition des fonds. Le montant du prêt accordé est en réalité diminué, dans la mesure où une partie de celui-ci doit être remboursée presque immédiatement au prêteur. Cette méthode permet toutefois de réduire le montant des versements, et a été mise en avant comme un moyen de maintenir des taux de remboursement de prêt élevés (ARMENDÀRIZ et MORDUCH, 2010).

Mais serait-elle en fait susceptible de décourager les investissements et de réduire les bénéfices pour les clients (et probablement la croissance de l'économie locale) ? Les emprunteurs pourraient-ils s'en sortir mieux s'ils avaient davantage de temps pour faire des investissements avant de commencer à rembourser leur prêt ? FIELD *et al.* (2013) ont conçu une RCT pour tester cette hypothèse, en se demandant si le contrat de microfinance « classique » empêchait d'investir dans des opportunités commerciales à fort rendement. Ils ont collaboré avec une organisation non gouvernementale (ONG) aidant les femmes dans des quartiers à faible revenu de Calcutta (Inde). Chaque client a reçu un prêt personnel d'un montant allant de 4 000 roupies indiennes (Indian National Rupee, INR) (90 $) à 10 000 INR (225 $), avec un montant de prêt modal de 8 000 INR.

Après la constitution de groupes et l'approbation des prêts (mais avant la mise à disposition des fonds), les groupes ont été répartis de manière aléatoire selon deux types de contrats. Dans le groupe de contrôle, 85 groupes ont été affectés au contrat de prêt régulier avec remboursement par versements fixes démarrant deux semaines après la mise à disposition des fonds. Dans le groupe de traitement, 84 groupes ont été affectés au contrat incluant une période de grâce de deux mois. Les autres caractéristiques du contrat demeuraient par ailleurs constantes. Le montant total des intérêts payés était identique, et une fois la période de remboursement lancée, tous les groupes effectuaient un versement toutes les deux semaines sur une période de 44 semaines lors d'une rencontre collective[9].

9. La durée des échéances étant plus longue pour le groupe de traitement (55 semaines contre 44 avant que la totalité du prêt ne soit remboursée), le taux d'intérêt effectif pour ce groupe était légèrement inférieur.

Trois ans plus tard, le nouveau contrat mis en place semblait être une réussite : les niveaux de profits des emprunteurs du groupe de traitement étaient supérieurs de 57 % en moyenne. Ils utilisaient également 81 % de capitaux en plus et prenaient plus de risques étant donné qu'ils investissaient davantage. Du point de vue de l'organisme prêteur, le problème était toutefois que les difficultés de remboursement avaient triplé : 52 semaines après la date à laquelle le prêt aurait dû être entièrement remboursé, 6 % du groupe de traitement n'avait pas effectué un remboursement intégral, contre moins de 2 % pour le groupe de contrôle. Ces difficultés de remboursement étaient d'une ampleur telle que le contrat n'était pas rentable en respectant les taux d'intérêt praticables.

Cette étude n'est pas une étude d'impact du type « est-ce que ça fonctionne ? ». Elle vise plutôt à analyser la nature des contrats et des contraintes, en comparant un type de contrat par rapport à l'autre. Au cours de l'étude, une estimation du rendement du capital a pu être faite (11 à 13 % par mois), suggérant que le fait d'avoir accès à davantage d'argent renforcerait le bien-être, mais ce n'était pas l'objectif principal de l'étude. La RCT permet surtout d'aborder une interrogation récurrente : pourquoi l'impact mesuré du microcrédit semble être si modeste (Banerjee *et al.*, 2015c) ? Les méthodes de crédit constituent-elles une partie du problème ? Peuvent-elles être améliorées ?

La RCT s'écarte des études de marché en ce qu'elle teste un produit réel, plutôt que de s'interroger sur les préférences pour tel ou tel scénario hypothétique. Une étude de marché pourrait en effet faire apparaître une préférence pour les versements différés, mais livrerait sûrement peu d'informations sur les conséquences en termes d'investissement, de résultats commerciaux et de remboursement des prêts. Il serait bien entendu possible de mener une expérimentation sans RCT, mais l'association des deux est une façon logique d'apporter un éclairage sur les comparaisons.

Les taux d'intérêt du microcrédit

Les exemples ci-dessus concernent les contrats de microcrédit, mais l'innovation la plus importante dans ce domaine a certainement été le choix d'augmenter les taux d'intérêt. Et ce n'était pas un changement évident. Les banques publiques ont été créées précisément pour octroyer des crédits subventionnés dans les régions pauvres parce que l'on pensait que les clients ne pourraient pas payer des taux d'intérêt élevés. Mais les leaders de la première heure officiant dans ce domaine ont fait face à la pression de devoir couvrir leurs frais de base, et il devenait impératif pour eux de générer des revenus d'intérêts. En se fondant seulement (ou presque) sur des indications fortuites, les microprêteurs sont partis du raisonnement que les ménages pauvres semblaient emprunter régulièrement auprès d'usuriers qui leur faisaient payer 5 ou 10 % d'intérêts par mois, et qu'appliquer des taux de 20 ou 30 % par an ne paraissait donc pas prohibitif. Selon le principe du caractère décroissant du rendement marginal du capital, ils ont également pensé que les entrepreneurs dépourvus de capitaux pourraient

obtenir un rendement élevé sur les premiers montants reçus (ARMENDÀRIZ et MORDUCH, 2010).

Leur devise est bientôt devenue : « les ménages pauvres ont besoin d'accès au crédit, mais pas de crédit bon marché. » Cette conclusion supposait implicitement que l'élasticité de la demande de prêt par rapport aux taux d'intérêt était de zéro (MORDUCH, 2000). Ils ont donc augmenté les taux d'intérêt. CULL *et al.* (2018) ont par exemple montré que, pour un échantillon de 1 330 organismes de microfinance entre 2005 et 2009, les taux d'intérêt moyens du microcrédit, corrigés de l'inflation, étaient de 25 % par an (21 % à la médiane). Ces taux d'intérêt ont permis aux organismes de microfinance de réduire la dépendance aux subventions, même s'ils n'étaient qu'un quart à fonctionner réellement sans subvention.

Les prêteurs se sont rassurés en se disant que les impacts sur la diffusion du dispositif en termes de taille de clientèle étaient limités. Mais l'hypothèse essentielle, à savoir que l'élasticité de la demande par rapport aux taux d'intérêt était nulle, n'avait pas été testée et pouvait difficilement l'être. Concernant les données disponibles, les défis étaient de plusieurs types : (1) les établissements de crédit changeant rarement les taux d'intérêt, il y avait donc peu de données à analyser ; (2) si les divers prêteurs appliquaient des taux d'intérêt différents, il y avait bien d'autres éléments qui différaient entre les établissements, et toute tentative de distinguer l'impact causal des taux d'intérêt en comparant les niveaux d'emprunt entre les établissements était vouée à l'échec ; (3) même si les taux d'intérêt pouvaient varier au sein des établissements, les différences s'appliquaient presque toujours à des produits différents destinés à des typologies de clients différentes. Là encore, il était difficile d'exploiter les variations. De leur côté, les études de marché ont toujours montré que les emprunteurs souhaitent bénéficier de crédits moins chers, mais il reste peu évident de savoir à quel point ils y sont sensibles.

DEHEJIA *et al.* (2012) ont fait une première tentative pour évaluer l'élasticité de la demande de prêt dans une configuration non randomisée appliquant la méthode des « différences de différences », en se servant d'une quasi-expérimentation (et non d'une RCT). *SafeSave*, un prêteur opérant dans les bidonvilles de Dacca (Bangladesh), appliquait un taux d'intérêt de 2 % par mois pour les prêts qu'il accordait, mais pensait que le taux devait être augmenté à 3 % pour couvrir ses frais[10]. Ainsi, lorsqu'il ouvrait de nouvelles agences, *SafeSave* y appliquait un taux de 3 %. Finalement, les agences plus anciennes ont été mises en conformité avec les nouvelles, offrant ainsi une possibilité d'observer les évolutions de la demande de prêt du fait du passage des taux d'intérêt de 2 à 3 % par mois dans les anciennes agences. La demande de prêt dans les nouvelles agences pouvait alors servir à contrôler les chocs macro-économiques et les conditions globales dans une configuration de différences de différences. La situation était

10. Pour information, j'étais à l'époque membre de la coopérative *SafeSave*, et j'appartenais à son comité de direction. Elle fait maintenant partie de l'ONG BRAC.

inhabituelle, dans la mesure où les prix étaient augmentés dans certaines agences et pas dans d'autres, les autres conditions restant identiques.

Contrairement à l'hypothèse d'une élasticité de zéro, Dehejia *et al.* (2012) estiment l'élasticité à long terme de plus de - 1,0. En d'autres termes, une augmentation de 10 % du taux d'intérêt entraîne une diminution de la demande supérieure à 10 %. Au lieu d'accroître considérablement les revenus, le relèvement du taux d'intérêt a légèrement réduit les revenus nets et l'emprunt. L'étude va directement à l'encontre des attentes, mettant aussi à mal un pilier important de la philosophie de la microfinance. Ce premier test rigoureux a montré que les clients étaient sensibles au taux d'intérêt, et qu'ils empruntaient moins lorsque le coût du prêt augmentait.

L'étude reposait sur des hypothèses fortes. L'hypothèse majeure était que le *timing* du passage de 2 à 3 % était effectivement aléatoire, c'est-à-dire indépendant de l'évolution de la demande dans le secteur, postulat qui reposait seulement sur les souvenirs du président de l'organisme prêteur. Il fallait aussi justifier de la comparabilité entre les agences afin d'interpréter les différences de différences en s'appuyant sur une démonstration de la similarité des tendances antérieures au changement. Par ailleurs, les résultats étaient fondés sur des données concernant les choix de 5 147 membres d'une institution particulière dans un ensemble d'agences situées seulement dans les bidonvilles de Dacca, ce qui n'était pas vraiment généralisable.

Et pourtant, le résultat a été pris au sérieux parce qu'il était plausible et venait contredire vivement les attentes des professionnels (des économistes notamment, qui partent du principe que les courbes de la demande ont tendance à être descendantes). L'étude a montré que les ménages, notamment les plus pauvres, *prenaient bien* en compte le coût du crédit, et que cela réduisait en conséquence la demande de prêts.

Une RCT a permis d'analyser un exemple plus large. Karlan et Zinman (2019) décrivent une étude réalisée dans un but identique sur la banque mexicaine de microfinance Banco Compartamos. Il s'agit du plus grand établissement de crédit d'Amérique latine, avec des millions d'emprunteurs, alors que ceux de *SafeSave* se comptent plutôt en milliers. Compartamos pèse parmi les organismes de microcrédit à vocation très commerciale, et pratique des taux d'intérêt aux alentours de 100 % par an (Rosenberg, 2009). La banque souhaitait réduire ses taux d'intérêt et Karlan et Zinman (2019) y ont vu l'opportunité d'évaluer l'élasticité des taux d'intérêt en persuadant Compartamos de réduire ses taux d'intérêt à différents niveaux dans différents lieux, en intégrant dans la démarche un traitement randomisé et des groupes de contrôle (tout comme l'étude *SafeSave* nécessitait également une certaine hétérogénéité entre les différentes agences).

Avec Compartamos, la randomisation a été opérée à l'échelle des agences en fonction de leur répartition à travers le Mexique. Quarante régions ont été affectées de façon aléatoire à un groupe « taux élevé », avec un coût du crédit inférieur d'environ 10 points de pourcentage aux taux d'intérêt existants. Quarante autres régions ont été affectées de façon aléatoire à un groupe « taux faible »,

avec un coût du crédit inférieur d'environ 20 points de pourcentage aux taux d'intérêt existants. Les élasticités ont été évaluées en comparant la demande de crédit entre les agences.

Comme dans l'étude de DEHEJIA *et al.* (2012), les emprunteurs se sont avérés être sensibles aux taux d'intérêt. KARLAN et ZINMAN (2019) ont estimé l'élasticité des taux d'intérêt après la première année à - 1,1, et à - 2,9 la troisième année. En outre, comme indiqué également dans DEHEJIA *et al.* (2012), ce changement n'a manifestement pas contribué à accroître les profits. Après le changement de taux d'intérêt, Compartamos avait davantage de clients, mais aussi davantage de frais.

Sans l'intervention des chercheurs, Compartamos aurait probablement réduit les taux d'intérêt partout en même temps, sans conserver de groupe de contrôle. Et si la banque avait volontairement choisi certaines agences pour commencer la démarche, il y aurait eu un risque de biais de sélection. La RCT a ainsi créé des variations utiles sur le plan analytique. L'utilisation de la randomisation par Karlan et Zinman a éliminé le souci de devoir comparer les comportements entre les agences. Elle a également permis d'éliminer la possibilité que la réduction des taux d'intérêt (et le montant de cette réduction) dans telle ou telle agence ait pu être induite par des contraintes locales. Au lieu de produire des estimations plausibles basées sur une série d'hypothèses (comme dans DEHEJIA *et al.*, 2012), les paramètres de la RCT évalués par KARLAN et ZINMAN (2019) sont transparents et mesurés avec justesse.

D'un autre côté, le cas de Compartamos est peu commun : le taux d'intérêt de départ était très élevé et le changement de politique consistait à réduire, plutôt qu'à augmenter, les taux d'intérêt. Comme c'était le cas avec l'étude sur *SafeSave*, les estimations obtenues avec Compartamos ne sont pas directement transposables à d'autres configurations. La conjonction des deux études peut néanmoins faire bouger les *a priori* au sens bayésien, et la démarche expérimentale qui sous-tend la RCT sur Compartamos a révélé certains points qu'il aurait été difficile de constater autrement[11].

Pauvreté, migration et argent mobile

La technologie est en train de transformer le paysage financier, en élargissant les possibles au-delà du microcrédit traditionnel, mais l'utilisation de technologies comme l'argent mobile (où on se sert des téléphones pour effectuer des paiements et gérer des porte-monnaie électroniques) relève fortement d'une auto-sélection. Le choix d'adopter une nouvelle technologie est renforcé par les politiques des fournisseurs visant à se concentrer sur les segments les plus lucratifs des marchés, ce qui signifie la plupart du temps que les ménages pauvres sont exclus de façon

11. La RCT ne répond pas à toutes les questions. Il semble que l'augmentation dans l'octroi de prêts découlait pour une large part de nouveaux emprunts (et non du remplacement d'autres sources), mais des questions subsistent quant aux impacts sur le bien-être et sur les risques de surendettement. En outre, la RCT est limitée dans ce qu'elle peut montrer du contexte et de l'hétérogénéité. Les résultats ne révèlent rien non plus quant aux questions éthiques liées au fait d'appliquer des taux d'intérêt relativement élevés à des emprunteurs pauvres (ROSENBERG, 2009).

disproportionnée. Le corollaire est que les ménages pauvres qui ont recours à ces moyens numériques sont plutôt rares. Comment peut-on alors évaluer les possibilités d'accès à la technologie des communautés pauvres ?

La population rurale du Bangladesh n'a cessé d'être attirée vers Dacca, principalement avec l'espoir de trouver un emploi dans l'industrie du prêt-à-porter, dont les usines, quelle que soit leur taille, suivent l'exemple de la Chine et exportent dans le monde entier. Les emplois sont souvent occupés par de jeunes travailleurs qui aident leurs familles restées dans les campagnes. Cette dynamique s'inspire du modèle de l'exode rural et de la croissance économique de Lewis (1954), le Bangladesh ayant un taux de croissance annuel d'environ 6 à 7 %. Mais qu'en est-il des ménages qui restent dans les zones rurales ? Une des questions est de savoir si la technologie peut aider les travailleurs qui migrent vers Dacca à envoyer de l'argent à leurs familles. La technologie contribue-t-elle à augmenter les transferts d'argent effectués par les travailleurs venus s'installer en zones urbaines (« migrants urbains ») au profit des familles rurales ? Cela peut-il être un moyen de réduire la pauvreté et les inégalités spatiales ?

Lee *et al.* (2021) utilisent une RCT basée sur un dispositif d'incitation aléatoire pour étudier de quelle manière l'accès à la banque mobile change la vie dans des communautés très pauvres du Bangladesh. Nous avons démarré avec un échantillon de ménages habitant dans les zones rurales du Nord-Ouest, définies comme étant « ultra-pauvres », ce groupe souffrant particulièrement de la pauvreté et de la faim pendant la période appelée « *monga* » (« saison maigre »). Ces foyers avaient participé à un programme mis en œuvre par une ONG locale pour aider leurs enfants majeurs à partir travailler dans des usines à Dacca. L'étude portait sur les deux parties de l'équation des transferts d'argent, c'est-à-dire à la fois sur les expéditeurs et sur les destinataires (ou bénéficiaires). À Dacca, nous avons suivi les « migrants urbains » originaires de la région Nord-Ouest. Et dans la région Nord-Ouest, nous avons suivi leurs familles élargies. Au sein du groupe de contrôle, seulement 11 % possédaient un compte en banque, et 20 % avaient régulièrement recours à l'argent mobile.

L'une des raisons expliquant le faible taux initial d'adoption de la technologie numérique était la barrière que représentaient les menus en langue anglaise sur l'interface téléphonique utilisée par les fournisseurs d'argent mobile. La principale intervention expérimentale, conçue par les chercheurs, consistait à former à l'utilisation de cette technologie des groupes affectés de façon aléatoire, à la fois dans le contexte urbain et dans le contexte rural. Les participants ont bénéficié d'une expérience pratique aux transferts d'argent, de traductions des menus et d'une aide pour l'ouverture de comptes (cette formation a coûté environ 12 $ par foyer). Le groupe de contrôle n'a reçu ni aide ni formation.

Le premier résultat constaté a été une nette augmentation du recours régulier à l'argent mobile, passant d'environ 20 % dans le groupe de contrôle à 70 % dans le groupe de traitement. Les transferts d'argent des « migrants urbains » à leurs familles en zone rurale ont augmenté de 30 % par rapport au groupe de contrôle, ces flux d'argent conduisant à une baisse de l'extrême pauvreté dans

la zone rurale. La consommation moyenne a augmenté de 7 % en moyenne par rapport au groupe de contrôle, et les bénéfices se sont avérés particulièrement importants pendant la « saison maigre ». Les migrants, de leur côté, faisaient davantage état d'une moins bonne santé physique et émotionnelle, cohérente avec la pression qu'ils s'imposaient pour travailler plus longtemps et augmenter ainsi les transferts d'argent permis par la technologie de banque mobile.

L'expérience objet de la RCT a réduit les barrières d'accès pour des groupes particulièrement exclus. Ceci se serait peut-être produit sans la RCT, mais l'intervention expérimentale a permis une comparaison claire avec un groupe de contrôle similaire à un moment donné de l'histoire où l'inférence causale était possible. En se focalisant sur le rapport migration-transfert d'argent, l'étude présente une voie alternative pour améliorer les conditions de vie en zones rurales. Les réponses classiques consistent à apporter des ressources dans les zones rurales par l'intermédiaire du microcrédit et des programmes dits de « progression » qui ont pour but d'accroître la productivité dans ces régions (voir la RCT réalisée par Bandiera *et al.*, 2017). Dans le cas relaté ici, le dispositif mis en place vise d'abord à aider les travailleurs ruraux à trouver des emplois plus rémunérateurs dans les villes, puis à favoriser un mécanisme de transfert des ressources des villes vers les zones rurales.

Même si l'on peut penser qu'il s'agit d'une évaluation de « ce qui fonctionne », cette étude doit plutôt être considérée comme une enquête sur les inégalités spatiales et une tentative pour définir si le partage des ressources au sein du foyer est limité par des coûts. L'intérêt de l'étude n'est pas de montrer que des Bangladais illettrés sont découragés par les menus en anglais nécessaires pour se servir de comptes bancaires mobiles : cela n'est pas surprenant et ne mériterait pas une étude aussi approfondie. L'intérêt était plutôt d'exploiter cette barrière (ainsi que le programme de formation visant à la surmonter) comme une façon d'introduire une variation entre ceux qui utilisent l'argent mobile et ceux qui ne s'en servent pas. Pour le formuler autrement, on peut dire que la barrière a été la clé pour constituer un groupe de traitement et un autre de contrôle (par le biais d'un modèle de promotion aléatoire), qui ont permis de cartographier les conséquences de l'accès à l'argent mobile pour les migrants et leurs familles. En définitive, l'étude n'a pas tant pour but de promouvoir une solution en particulier que de contribuer à la compréhension des voies pour sortir de la pauvreté rurale.

Défaillance du marché et biens privés

Les RCT sont par nature particulièrement utiles pour étudier des interventions isolées. Elles sont notamment bien adaptées aux études sur la fourniture de biens privés, comme le montrent les exemples développés ci-dessus. Dans le même ordre d'idées, les RCT sont beaucoup moins efficaces pour évaluer le rôle des biens publics et des changements macro-économiques (Hammer, 2017).

Certains reprochent aux RCT d'avoir focalisé l'attention de l'économie du développement sur la fourniture de biens privés, mais cette orientation prise par l'économie et les politiques de développement a vu le jour des décennies avant que les RCT n'arrivent sur le devant de la scène. L'évolution fondamentale de l'économie du développement vers les questions liées à la fourniture de biens privés remonte aux années 1970. Elle est apparue dans le contexte d'un mouvement plus large autour des problématiques du développement rural, des niveaux de pauvreté absolue, de la sous-alimentation, des taux de mortalité élevés et des niveaux d'instruction faibles, mouvement dont témoignent la littérature relative aux « besoins fondamentaux », les critiques du développement basé sur la croissance (par exemple, Chenery *et al.*, 1979), la réorientation de la Banque mondiale sous la présidence de Robert McNamara, la montée en puissance de l'économie de l'information au sein de l'économie du développement (Stiglitz et Weiss, 1981 ; Bardhan, 1984 ; Stiglitz, 1986) et l'intérêt accordé aux « biens tutélaires » (Musgrave, 2008)[12]. Les objectifs du millénaire pour le développement (OMD) et les objectifs de développement durable (ODD) établis et adoptés par l'ONU, dont les enjeux sont la pauvreté, la santé, l'éducation et les droits fondamentaux, renforcent encore cette tendance. L'une des raisons pour lesquelles les RCT se sont imposées est qu'elles sont justement particulièrement bien adaptées aux études sur la fourniture de biens et de services essentiels.

Comme le constatent Rodrik (2009) et Ravallion (2012), ceci attire l'attention vers des interventions modestes plutôt que les changements macro-économiques plus importants qui induisent la pauvreté, les inégalités et la croissance économique. Comme le soutiennent les critiques, concentrer l'attention sur les actions qui peuvent être étudiées par le biais des RCT entrave toute tentative pour mettre en œuvre des réformes systémiques dans les pays où les systèmes sont sérieusement défaillants, pervertis et injustes. Pour l'exprimer de façon plus crue, les RCT sont particulièrement adaptées pour étudier l'impact des « solutions sparadrap », et nous disposons en conséquence de nombreuses études en rapport avec celles-ci. Les RCT conviennent également bien à l'analyse des systèmes de distribution (la « problématique du dernier kilomètre »), plutôt qu'aux grandes priorités sectorielles (la « problématique du premier kilomètre » ?). Pour les détracteurs des RCT, nous devrions plutôt nous attaquer aux inégalités structurelles, aux conditions environnementales, aux déséquilibres politiques et aux infrastructures fragiles qui génèrent et reproduisent les maux que nos « solutions sparadrap » ne font que camoufler.

Cet argument des critiques est d'une importance fondamentale, et peut-être vaut-il mieux ne pas aller plus loin. Mais, en stoppant le débat ici, nous risquons de passer à côté d'une optique plus large, d'un conflit plus profond et de questions aussi importantes que sans réponse sur le rôle des sparadraps et des systèmes de distribution, des connaissances et du progrès.

12. L'enseignement est inclus ici en tant que « bien privé » car, au contraire des biens publics types, il est essentiellement « rival » et « exclusif ». Étant donné qu'il existe des effets externes manifestes pour la communauté dans son ensemble, l'enseignement doit peut-être plutôt être considéré comme un bien tutélaire.

Premièrement, cette réflexion montre explicitement que ces débats en apparence techniques et statistiques autour des méthodes appropriées pour garantir la validité interne et externe mériteraient plutôt une reconnaissance comme faisant partie du débat politique sur la portée et la nature de l'intervention même. Les discussions techniques peuvent se régler d'elles-mêmes grâce à l'innovation statistique et à l'amélioration des modèles de recherche, mais cela ne peut pas apaiser les tensions politiques plus fondamentales sur la portée des actions.

Deuxièmement, l'argument théorique en faveur des réformes systémiques est irréfutable. Le recul massif de la pauvreté enregistré au cours de ces dernières décennies à l'échelle mondiale, par exemple, est le fruit de transformations systémiques de grande ampleur, notamment en Asie (RAVALLION, 2012). Mais les changements systémiques ne sont pas toujours possibles, et des pans entiers de la population sont laissés pour compte. Étendre l'accès et la fourniture de services, notamment la mise à disposition de produits de base, reste une priorité essentielle pour les gouvernements, les organisations humanitaires et les fondations.

On pourrait raisonnablement penser que l'économie du développement *devrait* se concentrer davantage sur le contexte et sur les biens publics (HAMMER, 2017), sur les mesures macro-économiques et sur d'autres types de politiques, mais il est fallacieux d'affirmer que les RCT sont à l'origine des déséquilibres perçus. L'économie politique et l'histoire sont plus profondes, et il y a toujours et encore des raisons légitimes de continuer à améliorer la fourniture de biens et de services privés (même en l'absence de RCT). Les résultats des RCT ne déclencheront pas des révolutions, mais, appliquées cumulativement, elles peuvent créer des dispositions favorables pour améliorer les choses.

Conclusion

Les débats sur les RCT sont souvent peu satisfaisants. Ils ne permettent pas de faire précisément la distinction entre les différentes RCT et entre les types de sujets concernés. Les critiques sur les RCT sont pour la plupart convaincantes, tant pour des raisons philosophiques que techniques, et c'est à juste titre que leurs détracteurs avancent l'argument qu'elles ne constituent pas une source unique d'évaluations crédibles d'impact. D'autres approches sont également utiles et parfois meilleures. Nous avons besoin de davantage de descriptions, de données qualitatives, de *big data* et d'études reposant sur d'autres stratégies empiriques.

Et, dans le même temps, le débat ne fait pas ressortir ce que les RCT ont de vraiment innovant et passionnant. D'abord, les méthodes imparfaites ne sont pas toutes d'une imperfection égale. L'ajout de nouveaux outils comme les RCT permet d'élargir l'éventail des possibilités méthodologiques. Ensuite, le cadre lui-même a souvent besoin d'être bouleversé pour permettre d'y voir plus clair.

Les *randomistas* soulignent le rôle des RCT pour déterminer les mesures qui fonctionnent et celles qui ne réussissent pas. Pour ma part, je me suis concentré

ici sur les RCT qui permettent de déconstruire les structures économiques. La différence entre les RCT d'évaluation et les RCT exploratoires reflète l'écart entre l'approche qui consiste à analyser l'existant, et celle visant à confronter la théorie à la réalité pour créer de nouvelles possibilités de progrès. Associées à une démarche expérimentale, ces RCT créent des variations exogènes qui offrent une nouvelle manière d'évaluer le fonctionnement des marchés, des institutions et des processus majeurs.

Remerciements

Je remercie Isabelle Guérin, Florent Bédécarrats, François Roubaud, Tim Ogden et Martin Ravallion pour leurs commentaires, ainsi que Tim Ogden, Michael Kremer, Lant Pritchett et les participants à la conférence sur les RCT en développement organisée par l'Agence française de développement les 19 et 20 mars 2019, pour toutes les discussions que nous avons pu avoir ensemble. J'assume la pleine responsabilité de mes opinions et erreurs.

Chapitre 4

Les expérimentations aléatoires dans l'économie du développement, leurs détracteurs et leur évolution

Timothy Ogden

Introduction

Pascaline Dupas voulait simplement aider certaines des personnes les plus pauvres au monde (cette histoire est tirée d'Ogden, 2017). Mais elle a découvert – à l'instar de beaucoup d'autres, malheureusement, et, plus malheureusement encore, à l'inverse de beaucoup d'autres – que les structures éducatives dédiées à la formation des étudiants riches dans les pays riches ne transmettent que très peu de compétences pratiques pour apprendre à vivre au sein de ménages pauvres dans les pays en développement ou à les « aider ». Autrement dit, elle n'a pas trouvé d'emploi dans une ONG internationale. Dépitée, elle a dû se contenter de son second choix : une bourse de recherche à Harvard.

Un poste d'assistante de recherche au Kenya lui a été proposé dans le cadre d'une expérimentation randomisée de terrain. Elle a très vite renoncé à sa bourse pour se rapprocher littéralement de son objectif initial, à savoir aider les populations pauvres des pays en développement. Pendant son séjour au Kenya, elle s'est liée d'amitié avec une jeune mère – et a vu cette femme se démener pour se procurer des médicaments et soigner son enfant atteint de paludisme. Une fois encore, comme beaucoup d'autres, elle a commencé à se demander pourquoi. Pourquoi cette femme avait-elle tant de mal à se procurer des médicaments, à mettre quelques dollars de côté, pour éviter en premier lieu que son bébé ne contracte le paludisme ? Lorsque, de retour en France, ses amis lui ont demandé ce qu'ils pouvaient faire pour aider, à qui ils pouvaient donner de l'argent pour faire une différence, elle a suggéré d'acheter des moustiquaires. Quand elle s'est

rendu compte qu'il n'y avait que peu d'organisations caritatives, voire aucune, qui utilisaient leurs dons pour acheter des moustiquaires, elle et quelques amis (qui ont également fini par devenir des économistes professionnels) ont créé une organisation caritative avec cet objectif très précis.

Lorsque cette organisation caritative a été critiquée pour avoir donné des moustiquaires – des critiques qui, inspirées des théories économiques classiques, prétendent que les gens dévalorisent les choses qui sont gratuites et que la gratuité fausse les marchés –, elle n'a ni admis ni réfuté ces critiques. Elle a décidé que la meilleure chose à faire était de tester ces critiques pour voir si elles étaient fondées. Elle a donc conçu une expérimentation randomisée pour faire varier les prix et la mise en œuvre de subventions pour les moustiquaires, dans le but de déterminer le mode optimal de distribution de moustiquaires. L'expérimentation s'est révélée très fructueuse sur les plans quantitatif et qualitatif. L'article a été cité plus de 570 fois selon Google Scholar et figure parmi les 5 % les plus cités de tous les articles de recherche selon Altmetric. Il a fortement influencé les recommandations politiques sur la distribution de moustiquaires. GiveWell, s'appuyant sur cette étude, notamment en termes de mise en œuvre de programmes de distribution de moustiquaires, a consacré 100 millions de dollars à la distribution gratuite de moustiquaires[1]. Dans l'ensemble, on estime que la distribution de moustiquaires a permis d'éviter 450 millions de cas de paludisme et 4 millions de décès (Bhatt *et al.*, 2015 ; Glennerster, 2016). Si, dans les années 2000, les effets des subventions et de la distribution gratuite de produits de santé ont fait l'objet de nombreux débats – débats qui, pendant des années, ont opposé des théories concurrentes et des principes issus d'études non expérimentales –, ces débats ont aujourd'hui largement disparu.

Cette histoire synthétise l'essentiel des débats sur les évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) en économie du développement : motivations, impact, validité interne et externe, portée, théories du changement, éthique et hypocrisie. Les dits *randomistas* sont connus pour critiquer le recours aux anecdotes, entre autres, et pourtant il s'agit là d'une anecdote. Les *randomistas* exigent des preuves empiriques pour les allégations de causalité, et pourtant il n'existe probablement pas de preuves randomisées démontrant que les recherches en question ont amené qui que ce soit à changer d'avis ou de pratique. Les détracteurs affirment que les RCT n'ont « rien de spécial », et pourtant des preuves comme celle-ci semblent avoir été plus convaincantes pour des non-économistes que les allégations théoriques ou scientifiques produites par d'autres méthodes[2]. Les détracteurs prétendent que les RCT limitent les questions susceptibles d'être posées et d'obtenir une réponse à des questions étroites et secondaires, et pourtant sauver 4 millions de vies est bien plus important que

1. Remarque : je suis président de GiveWell ; ces données proviennent d'entretiens avec le personnel de GiveWell.

2. Dans Ogden (2017 : 201-204), Frank DeGiovanni déclare que la fondation Ford a financé des RCT de programmes « *Targeting the Ultra Poor* » parce que la fondation avait estimé que les preuves empiriques issues des RCT étaient plus convaincantes pour les décideurs politiques (une autre anecdote !).

l'impact dont peuvent se targuer de nombreux économistes concernant leurs études sur les « grandes » questions. Les détracteurs dénoncent l'inutilité des évaluations « externes » qui n'influencent pas la politique ou la pratique, et pourtant cette RCT procède directement d'une question de pratique posée par un des dirigeants d'une organisation caritative.

Le débat sur les RCT remonte aussi loin que les RCT elles-mêmes. Ce volume est au moins le quatrième – après Cohen et Easterly (2010), Teele (2014) et Ogden (2017) – à rassembler des points de vue contradictoires sur le rôle des expérimentations randomisées de terrain dans le champ des sciences sociales, de l'économie du développement et des politiques de développement. Une étude de l'histoire de ces débats donnerait raison au dicton selon lequel la science progresse « funérailles après funérailles[3] », puisque personne ne change d'avis. J'aborde le sujet avec une certaine appréhension. Ma contribution – j'ose espérer qu'elle en sera une – consiste à tenter de résumer et de systématiser les critiques les plus importantes, d'examiner les difficultés à apporter une réponse significative et cohérente à ces critiques et, enfin, de discuter de l'évolution de la pratique des RCT. Il est impossible de dire si cette évolution s'est faite ou non en réponse aux critiques, mais le fait est que l'évolution *a* répondu aux critiques. Je conclus par un modèle de réflexion sur l'évolution des RCT et des méthodes empiriques et expérimentales, leur situation présente et leur avenir probable.

Les critiques des RCT

Je souhaite ici donner un aperçu de ce que je perçois comme étant les principales critiques de l'utilisation des RCT en économie (et en sciences sociales) afin de donner un semblant d'ordre aux réponses. Je regroupe les critiques en sept catégories :

– les critiques « rien de magique » (*Nothing Magic*) ;
– les critiques « boîte noire » (*Black Box*) ;
– les critiques « validité externe » (*External Validity*) ;
– les critiques « signification négligeable » (*Trivial Significance*) ;
– les critiques « saucisse politique » (*Policy Sausage*) ;
– les critiques « éthiques » (*Ethics*) ;
– les critiques « trop » (*Too Much*).

Il y a d'autres critiques et d'autres nuances de ces critiques, que je ne relève pas ici. Je ne doute pas que certains détracteurs s'opposeront à la façon dont je qualifie leurs arguments, mais il faut bien commencer quelque part.

3. N.D.E. : l'expression renvoie à l'idée selon laquelle certaines idées scientifiques n'avancent pas parce qu'elles parviennent à convaincre leurs opposants, mais plutôt parce que les opposants décèdent et leurs successeurs grandissent en étant familiarisés avec cette idée scientifique.

Les critiques « rien de magique »

Cette critique est ainsi nommée parce que son expression la plus directe est « Il n'y a rien de magique dans les RCT ». Cette critique répond à l'idée que les RCT « trônent au sommet d'une hiérarchie de méthodes » (Ravallion, chap. 1, ce volume) pour estimer l'impact causal.

La principale version de la critique « rien de magique » est que la randomisation ne produit pas nécessairement une estimation d'impact moins biaisée que d'autres méthodes. Deaton et Cartwright (2018) élaborent la discussion la plus complète qui soit sur la principale forme de cette critique. Cook (2018) détaille 26 hypothèses requises pour croire qu'une RCT produit de fait une estimation non biaisée. Cette critique évoque souvent ou prolonge les débats sur les RCT qui remontent jusqu'à la critique de Fisher par Student (1938).

Une autre version de la critique « rien de magique » est que les expériences de terrain en économie ne se conforment pas à la norme en double aveugle des RCT dans la pratique médicale – et son appellation pourrait donc être complétée ainsi : « rien de magique dans les RCT en économie du développement ». L'incapacité à pratiquer des essais en double aveugle, voire des essais à l'aveugle, signifie que les RCT en sciences sociales ne permettent généralement pas de réduire l'une des principales sources de biais attendues.

Une troisième version de la critique dit que, même si les RCT limitent les degrés de liberté, rien n'est éliminé. Les RCT doivent donc faire l'objet d'analyses tout aussi minutieuses que d'autres méthodes. Ioannidis (2018) résume les résultats de plusieurs examens de RCT qui trouvent des preuves significatives de biais chez des RCT publiées dans plusieurs disciplines. Young (2019) constate que la majorité des RCT publiées échouent aux comparaisons multiples et aux tests rétroactifs de significativité. Kaplan et Irvin (2015) étudient les résultats d'essais médicaux utilisant des RCT et constatent que le nombre de résultats « sans effet » a nettement augmenté lorsque les chercheurs ont dû établir un plan de pré-analyse expliquant précisément comment ils comptaient évaluer les données recueillies avant de lancer l'expérience. En outre, les RCT seraient tout aussi vulnérables aux faux positifs, aux faux négatifs et aux erreurs d'amplitude que n'importe quelle autre méthode de recherche (Gelman, 2018).

Les critiques « boîte noire »

Il est un autre groupe de critiques, étroitement lié aux critiques « Rien de magique », qui postule que ce que l'on peut apprendre de la plupart des RCT se limite à savoir si une intervention a « fonctionné », et non pas *pourquoi* elle a fonctionné. Une RCT n'éclaire pas nécessairement le véritable mécanisme causal, même lorsqu'une relation causale est établie de manière convaincante. Un exemple utile est l'interview de James J. Choi et Dean Karlan au sujet d'une RCT qu'ils ont conduite, où les deux (et le responsable de la mise en œuvre) ne s'accordent pas sur la cause profonde de leurs résultats (Dubner, 2018). Les RCT qui ne trouvent aucun effet peuvent être pires encore. Dans bien des cas, il

est impossible de déterminer si le résultat nul est dû à un traitement inefficace ou à une mise en œuvre inefficace du traitement. L'interprétation de résultats nuls est souvent très floue, même parmi les partisans des RCT (Evans, 2016).

Une variante de la critique « boîte noire » est la critique « athéorie ». Une RCT qui n'est pas fondée sur la théorie peut être très difficile à interpréter, que le résultat puisse être distingué de zéro ou non. Mais, s'il y a une théorie bien fondée qui sous-tend la RCT, les avantages de la randomisation peuvent être assez limités. Il est possible de concevoir des approches alternatives (et plus simples à mettre en œuvre) pour tester une théorie clairement définie.

Les critiques « validité externe »

Les critiques « validité externe » insistent sur le fait que chaque RCT est ancrée dans un contexte très spécifique. Cela couvre des aspects tels que l'opérateur d'une intervention (souvent une ONG), le personnel engagé par cette ONG, la culture et les coutumes locales et régionales, la technique d'enquête, la manière spécifique dont les questions sont posées, et même la météo. Ainsi, si les résultats d'une RCT particulière peuvent nous en apprendre beaucoup sur l'impact d'un programme particulier dans un endroit particulier à un moment particulier, ils ne nous apprennent pas grand-chose sur le résultat de l'exécution d'un programme strictement identique dans un contexte et à un moment différents. La critique « validité externe » est traitée en détail dans le livre de Nancy Cartwright et Jeremy Hardie intitulé *Evidence-Based Policy*. Cartwright et Deaton (2018) contribuent également à cette critique dans leurs articles, tout comme ici Pritchett (chap. 2, ce volume) et Ravallion (chap. 1, ce volume).

Les critiques « signification négligeable »

Je donne à cette critique le nom de « signification négligeable » (*Trivial Significance*) pour la distinguer de l'usage courant du terme « significativité » dans les discussions statistiques. J'utilise ici le terme comme synonyme de « importance relative » dans le vocabulaire des affaires et de la comptabilité. La critique « signification négligeable » n'est pas d'ordre statistique et ne porte pas sur la taille de l'effet relatif, mais sur la taille de l'effet absolu (vraiment absolu, parfois) : il importe de déterminer si les programmes et les politiques du mouvement des RCT s'intéressent à ce sujet.

Les critiques peuvent revêtir plusieurs formes différentes, mais toutes partagent le même propos fondamental : les programmes et les projets mesurés et mesurables par des RCT produisent des changements qui, même lorsqu'ils sont « réussis », ne sont pas assez importants pour faire une différence entre la pauvreté et la prospérité, compte tenu de l'ampleur du problème de la pauvreté mondiale. Ces critiques peuvent provenir d'une perspective macro-économique (les choses qui « comptent vraiment » sont des choix de niveau macro-économique, comme la politique commerciale qui ne peut pas être randomisée) (Ravallion, chap. 1, ce volume), d'une perspective systémique (une RCT sur la hausse des

taux de vaccination n'améliore pas le système de santé et peut en fait entraver le développement du système) (Garchitorena *et al.*, chap. 5, ce volume) ou d'une perspective d'économie politique (les RCT ne peuvent pas permettre de déterminer si la croissance peut être favorisée par des investissements dans les infrastructures de transport, dans les systèmes de santé ou dans les systèmes d'éducation) (Hammer, 2014).

Les critiques « saucisse politique »

Les critiques « saucisse politique » (*Policy Sausage*) sont principalement associées à Pritchett. La version simplifiée est que les politiques (qu'elles soient gouvernementales ou propres aux ONG) sont le fruit d'actions complexes et opaques influencées par la politique, les capacités, les contraintes de ressources, l'histoire et de nombreux autres facteurs. L'élaboration de politiques est comme la fabrication de saucisses. L'évaluation d'impact – et la recherche académique indépendante en général – ne joue qu'un rôle mineur dans la saucisse politique, surtout si l'évaluation d'impact vient de l'extérieur de l'organisation. Ainsi, l'effort consacré à une RCT est probablement gaspillé, car il ne sera d'aucun effet sur ce processus complexe. Bédécarrats *et al.* (2019b) notent le nombre très limité de programmes évalués par le biais d'une RCT qui semblent avoir été déployés à plus grande échelle.

Pritchett et d'autres soutiennent que le processus de changement politique ou organisationnel est totalement distinct de la démarche de création de savoir. Le pont entre les deux n'est pas issu de volontés politiques, mais implique un travail minutieux au sein des bureaucraties, des machines politiques et des organisations. Plus spécifiquement, Pritchett suggère que le modèle d'adoption des politiques, supposé par les *randomistas*, est « incroyablement préhistorique » (*unbelievably Cro-Magnon*) (Ogden, 2017).

Les critiques « éthique »

Depuis que les RCT existent, elles sont critiquées sur le caractère non éthique de la méthode. Cette critique se présente sous deux grandes formes. L'une d'elles est que l'expérimentation sur des êtres humains, en particulier – et surtout – sur des personnes issues de communautés pauvres (comme c'est nécessairement le cas dans les RCT sur l'économie du développement), est intrinsèquement contraire à l'éthique. Meyer *et al.* (2019) font observer que cette intuition morale, qui incrimine l'expérimentation, est très répandue, au moins au sein de la population américaine.

Historiquement, la communauté de la recherche médicale a abordé cette aversion morale quant à l'expérimentation et au refus de traitement par le biais du concept d'équipoise – ne pratiquer une expérimentation que quand il y a une incertitude raisonnable concernant les bénéfices (ou les bénéfices relatifs) d'un traitement (Freedman, 1987). Dans le contexte de l'économie du développement, Abramowicz et Szafarz (chap. 10, ce volume) font valoir que le concept

d'équipoise a été trop rapidement escamoté à l'aune de la pauvreté et des privations de nombreux participants aux études RCT et qu'il est souvent tout bonnement ignoré par les partisans de la méthode.

Les critiques « trop »

La critique « trop » (*Too Much*) signifie que, même si les RCT ont leurs avantages par rapport à d'autres solutions, ces avantages ne justifient pas les coûts qu'elles imposent en termes de temps, d'argent, d'opportunité ou de « talent ». Ravallion (2009a ; chap. 1, ce volume), par exemple, soutient qu'il y a trop de RCT, qu'elles évincent les évaluations de programmes qui ne sont pas adaptés à la randomisation, et sous-entend qu'elles occupent trop de place dans les revues spécialisées. Pritchett estime que les « principaux fondateurs du mouvement sont tous des génies » qui devraient laisser les RCT aux « doctorants en santé publique de l'État du Kansas » (Ogden, 2017 : 144). Deaton suggère qu'elles conviennent mieux aux consultants des gouvernements pour régler des différends politiques, plutôt que de venir nourrir les connaissances académiques (Ogden, 2017). D'autres déplorent que les délais de réalisation et d'analyse des RCT soient trop longs par comparaison avec d'autres méthodes, même imparfaites. Par ailleurs, la méthode exige généralement que les organisations qui mettent en œuvre un programme fassent perdurer l'intervention pendant toute la période de l'évaluation, indépendamment du retour d'information, ce qui induit des coûts d'opportunité élevés (Whittle, 2011).

Le défi de répondre aux critiques

Les promoteurs des RCT ne sont pas restés passifs sous le feu de ces critiques, loin s'en faut. Comme indiqué, on ne compte plus les ouvrages et autres documents, blogs, interviews, dossiers ou livres qui énumèrent des arguments en faveur des RCT et répondent à des critiques ou à des détracteurs en particulier. Mais, au fil du temps, les réponses se sont faites plus rares. Ces dernières années, les promoteurs des RCT ont apparemment commencé à refuser tout dialogue avec leurs détracteurs (bien que, comme je le soutiens plus loin dans ce chapitre, ils aient accordé une certaine attention aux critiques, même indirectement).

L'une des raisons est peut-être que, sur le plan rhétorique, les défenseurs apparaissent quelque peu désavantagés par rapport aux détracteurs – et la situation réitère peut-être certains des premiers échanges dans les débats, lorsque c'étaient les partisans des RCT qui critiquaient les méthodes établies. En bref, le détracteur a simplement à trouver quelques exemples d'un problème particulier, alors que la réponse doit défendre un système établi, mais en constante évolution.

Qu'est-ce qu'un *randomista* ?

Une illustration spécifique de cette difficulté réside dans la question fondamentale de la définition du *randomista*. De nombreuses critiques sont fondées sur les convictions des *randomistas* – mais il n'y a pas de manifeste ou de profession de foi disant qui peut faire partie du club. Lant Pritchett a décrit le mouvement RCT comme un mouvement religieux, et les émotions qu'il suscite semblent assurément valider cette comparaison. Pour reprendre la définition de Stackhouse, une religion ou un mouvement est « une vision globale du monde ou une "vision morale métaphysique" qui est acceptée comme contraignante parce qu'elle est considérée comme fondamentalement vraie et juste, même si toutes ses dimensions ne peuvent être ni pleinement confirmées ni réfutées » (STACKHOUSE, 2007 : 7). À première vue, en particulier pour qui a assisté à de nombreux débats publics ou discussions sur la valeur des RCT, cette description peut tout à fait convenir aux deux parties du débat. Mais des discussions approfondies avec les personnes concernées érodent rapidement le sentiment d'une « vision globale du monde » commune au regard de la valeur, des avantages et de l'applicabilité des RCT. Les entretiens que j'ai menés pour *Experimental Conversations* (OGDEN, 2017), avec dix des principaux praticiens des RCT, apportent de nombreuses preuves empiriques de l'hétérogénéité des croyances, ne serait-ce que parmi ce petit échantillon de partisans des RCT.

En l'absence de profession de foi énonçant les croyances fondamentales du *randomista*, les détracteurs ont tendance à s'appuyer sur une construction rhétorique qui peut être rendue ainsi : « Puisque l'individu X a dit Y en t1, le groupe a tort en t2 ». L'un des arguments favoris des détracteurs est l'axiome indémodable de BANERJEE (2006) : « Les expérimentations randomisées […] sont simplement la meilleure façon d'évaluer l'impact d'un programme ». À leur crédit, DEATON et CARTWRIGHT (2018) puisent dans un plus grand nombre de citations et mettent notamment en exergue des déclarations tirées des travaux d'Abdul Latif Jameel Poverty Action Lab (J-PAL). Celles-ci constituent certes un bien meilleur point de départ pour une critique, tout en demeurant limitées, à moins d'invoquer la culpabilité collective ou la culpabilité par association. Le problème n'est pas que ces déclarations soient dénuées de sens ou que personne ne les croie en particulier, mais qu'il est très difficile de déterminer qui exactement souscrirait à telle ou telle déclaration, et si ceux qui y souscrivent ou n'y souscrivent pas peuvent raisonnablement être rangés à l'intérieur ou à l'extérieur du cercle des *randomistas*.

La tentative la plus proche d'une définition du *randomista* semble être celle de Ravallion (chap. 1, ce volume), qui se contente de définir un *randomista* comme quelqu'un qui croit que « les RCT trônent au sommet d'une hiérarchie de méthodes, qui pense que les RCT sont "l'étalon-or" des évaluations d'impact – l'approche la plus "scientifique" et "rigoureuse", garantissant une évaluation d'impact essentiellement athéorique et sans hypothèse, mais fiable ». Mais, même cette déclaration laisse beaucoup à désirer, puisqu'elle manque de la précision nécessaire pour définir un groupe. Si certains, comme IMBENS (2018) (notez toutefois qu'Imbens est avant tout un économétricien et non un économiste du développement), souscrivent à l'idée qu'il existe une hiérarchie, avec les RCT

« au sommet », suffit-il de croire qu'il y a une hiérarchie ou faut-il ménager un certain espace entre les RCT et d'autres méthodes ? Comment cet espace serait-il mesuré ? Même la formulation, standard en économie – selon laquelle « *ceteris paribus*, les RCT sont une meilleure méthode pour l'évaluation d'impact » –, laisse des abîmes de termes indéfinis, avec une hétérogénéité significative lorsqu'elle est employée entre *randomistas*, et probablement moins d'espace entre un prétendu *randomista* et un détracteur. Par exemple, considérez les citations suivantes (parfois légèrement modifiées pour supprimer les « non-dits » évidents[4]) et essayez de déterminer lesquelles sont d'un *randomista* et lesquelles sont d'un détracteur.

1. « Les meilleures méthodes à utiliser et leurs combinaisons possibles dépendent de la question exacte en jeu, du type d'hypothèses de base qui peuvent être raisonnablement employées et du coût des différents types d'erreurs » ;

2. « Nous ne devrions pas encourager ou décourager l'emploi d'un outil particulier juste pour ce qu'il est. Nous devrions encourager les étudiants à poser une question intéressante et à employer le bon outil pour y répondre. Point » (Karlan *in* OGDEN, 2017 : 86) ;

3. « [De] très bonnes données descriptives qui attirent l'attention des gens sur quelque chose qui ne les avait jusque-là pas intéressés ont changé l'opinion des gens en matière de politique, autant que n'importe quelle expérience » (McKensie *in* OGDEN, 2017 : 134) ;

4. « La nouveauté […] a conduit à une surenchère sur les RCT, comme ces déclarations stupides affirmant que tout doit être randomisé pour être évalué, ou les gens qui disent ne croire à aucune preuve issue de données d'observationnelles » (Yang *in* OGDEN, 2017 : 238) ;

5. « [Aucune approche] ne fait disparaître tous les problèmes, pas plus les RCT qu'une autre. Je ne pense pas [qu'il faille croire] que les RCT sont magiques » (Kremer *in* OGDEN, 2017 : 19) ;

6. « [C]'est en partie ainsi que le mouvement d'évaluation randomisée a été vendu aux décideurs politiques : "Vous allez avoir des réponses". Je ne crois pas que ce soit là ce que nous allons avoir. J'ai le sentiment que nous allons voir des évaluations partout dans le monde. Si je devais choisir, je […] dirais que nous devrions insuffler plus d'énergie dans les grandes choses que dans les petites » (Blattman *in* OGDEN, 2017 : 227) ;

7. « Les organisations devraient pouvoir puiser dans différents domaines pour répondre aux questions pertinentes […] Je constate de nombreux recoupements entre différentes formes d'identification causale […] Je ne pense pas que vous […] vous intéressiez aux évaluations randomisées. Je ne pense pas que cela fasse sens » (Glennerster *in* OGDEN, 2017 : 200) ;

8. « Une évaluation d'impact devrait permettre de déterminer pourquoi quelque chose fonctionne, et non pas simplement *si* cela fonctionne. Il ne faudrait pas

4. Voir dans l'annexe de ce chapitre les citations originales non modifiées si vous souhaitez juger de la pertinence de ces modifications.

faire [de RCT] si elles ne peuvent fournir aucune connaissance généralisable sur la question du "pourquoi" » (Gugerty et Karlan, 2018 : 45) ;

9. « Il y a de nombreux exemples banals et inutiles d'études appliquant chaque méthode spécifique » (McKensie, 2018).

Si n'importe laquelle de ces citations peut être écartée ou contestée d'une manière ou d'une autre, il n'en demeure pas moins que de nombreux prétendus *randomistas* expriment à maintes reprises le sentiment que les RCT sont un « outil dans la boîte à outils » de l'économie moderne, qu'il existe de nombreux autres outils utiles, que les RCT ne conviennent pas à toutes les questions dignes d'intérêt et que d'autres outils analytiques sont tout aussi utiles et crédibles. De plus, la plupart des *randomistas*, si ce n'est la totalité, utilisent et publient d'autres méthodologies. Comme l'a souligné McKenzie (2019), les papiers les plus cités des trois *randomistas* les plus connus – Banerjee, Kremer et Duflo – ne sont pas des RCT.

La seconde moitié de la définition de Ravallion (« l'approche la plus "scientifique" ou "rigoureuse", promettant de livrer une évaluation d'impact largement athéorique et sans hypothèse, mais fiable ») serait encore plus difficile à faire accepter par un nombre significatif d'économistes pratiquant des RCT. En économie, le débat sur le bon ordonnancement de l'analyse des données et de la théorie, qui dure depuis des décennies, est aussi insoluble que le débat sur les RCT (Cherrier, 2019). La qualification « largement athéorique et sans hypothèse » pourrait à elle seule inspirer (et a inspiré) des pages et des pages de débat. C'est ce fait même qui a conduit aux *Experimental Conversations*, où j'ai décidé que la seule chose cohérente à faire était d'interviewer de nombreux *randomistas* et quasi-*randomistas* pour les entendre explorer les nuances de leurs croyances concernant les pratiques véritablement mises en œuvre dans la conduite et l'interprétation des études, et non pas seulement la métaphysique des méthodologies.

Le but de cette discussion n'est pas de conclure que le terme « *randomista* » est si mal défini et indéfinissable qu'il est pratiquement inutile, bien que je pense que ce soit le cas, mais d'illustrer pourquoi il est si difficile de répondre intelligemment aux critiques (et peut-être d'expliquer pourquoi les *randomistas* ont généralement cessé de répondre directement aux critiques, comme en témoigne le fait qu'aucun n'ait accepté de participer à ce volume). Quelle que soit la réponse, elle doit nécessairement procéder d'une personne qui sera probablement d'accord avec certaines parties au moins d'une critique particulière – et qui, finalement, ne parlera que pour elle-même. Dans le même temps, toute personne qui, comme je le fais dans ce chapitre, tente de défendre les RCT doit assumer une certaine responsabilité pour répondre de toute déclaration formulée au nom des RCT, qu'elle soit improvisée, imprudente, malavisée ou erronée. Il n'est donc pas surprenant que peu de gens apprécient, à ce stade tardif des discussions en cours, d'avoir l'opportunité de répondre à une critique de *randomistas* mal définis en général en faisant spécifiquement état de leurs croyances personnelles. À quoi cela mènerait ?

L'argumentation derrière les arguments

Comme je ne suis pas un *randomista*, dans le sens où je ne fais pas carrière dans la pratique des RCT, je peux difficilement m'arroger le droit de répondre aux critiques ou aux détracteurs au nom des *randomistas*. Cela étant, dans la prochaine section, je me servirai des actions (études et autres activités professionnelles) de divers *randomistas* nominaux pour illustrer l'évolution du mouvement, en démontrant que celle-ci a au moins eu pour effet d'émousser bon nombre des critiques. Mais, auparavant, je voudrais évoquer un autre problème qui ne facilite guère les réponses directes et constructives aux critiques.

Les différends entre les *randomistas* et leurs détracteurs mettent souvent trop en avant les particularités de la méthodologie, masquant ainsi le conflit sous-jacent bien plus important. Ce conflit porte sur les théories du changement. Les débats autour des théories du changement – les idées sur la façon dont le monde change – ne sont pas spécialement propres au moment présent en économie du développement. C'est en effet le fondement même de l'économie du développement (et d'une grande partie des autres sciences sociales) : comment se fait-il que des pays pauvres deviennent plus riches (ou pourquoi les pays pauvres restent-ils pauvres) ?

Les théories du changement ont toujours fait l'objet de controverses au sein de la profession économique. La plupart des textes fondateurs de l'économie peuvent être vus comme des manifestes sur les théories du changement. L'économie du développement a sa propre théorie spécifique des conflits liés au changement : par exemple Sachs (2005) contre Easterly (2007) ; « les institutions comptent » d'Acemoglu et Robinson (2006) ; la politique industrielle de Rodrik (2008b) ; les arguments anti-aide de Deaton (2013b).

Bien que je me méfie de la réduction des théories du changement à des résumés ou à des points dans un graphique, je trouve néanmoins utile, dans le contexte de cette discussion, de réfléchir aux théories du changement concurrentes selon trois grands axes :

– la valeur des petits changements par rapport aux grands ;

– la valeur des connaissances locales par rapport à l'expertise technocratique ;

– le rôle des actions individuelles par rapport aux actions collectives institutionnelles.

Ces axes ne sont pas totalement indépendants les uns des autres. Quelqu'un qui croit fermement à la valeur des grands changements a évidemment de grandes chances de priser l'expertise technocratique et le rôle des institutions. En pratique, il existe des variations significatives au sein du mouvement RCT et parmi les détracteurs des RCT, de sorte que, dans certains cas, il y a plus de points communs entre un partisan particulier des RCT et un détracteur particulier qu'entre deux détracteurs différents – ce qui pose, une fois encore, le problème de la définition du *randomista*.

Chacune des critiques des RCT mentionnées ci-dessus est sous-tendue par une théorie du changement qui diffère de celle des partisans des RCT selon au moins un des trois axes. Suite aux nombreuses conversations que j'ai eues avec des *randomistas* et des détracteurs, mon impression est que les membres du mouvement

RCT ont tendance à croire que de petits changements peuvent avoir une grande incidence, que l'expertise technocratique est extrêmement précieuse et que les individus au sein des institutions comptent autant que les institutions elles-mêmes. Les détracteurs qui invoquent la critique de la « signification négligeable », en revanche, s'accordent généralement sur la valeur de l'expertise technocratique, mais pas sur la valeur des petits changements et le rôle des institutions.

Cette différence est importante, car elle influence la façon dont on évalue la qualité et, surtout, l'utilité des preuves empiriques. Si vous pensez que l'amélioration du monde passe par une accumulation de petits changements, alors vous placez beaucoup plus bas la barre de la validité externe. Vous n'avez pas à avoir la conviction qu'un programme produira exactement le même impact ou un impact proche de celui observé dans un contexte différent pour qu'il vaille la peine d'être répliqué ailleurs. Il vous suffit de croire qu'il constitue un bon point de départ pour mener ailleurs une autre petite expérience en l'ajustant au fur et à mesure. La distinction entre un *randomista* – avec cette théorie du changement – et un afficionado des boucles de rétroaction comme Dennis Whittle porte simplement sur la vitesse d'itération et sur la valeur à accorder aux différents types de rétroaction. Bien sûr, ce même *randomista* regardant une RCT qualifiée d'athéorique par un critique serait totalement désorienté – il y a bien une théorie, mais peut-être pas de modèle structurel permettant d'estimer en toute connaissance de cause l'impact d'un contexte à l'autre. Parallèlement, un *randomista* ayant une théorie du changement légèrement différente, qui accorderait plus de valeur aux institutions, refusera de défendre les RCT à petite échelle des ONG et plaidera pour un surcroît d'expérimentation à grande échelle (par exemple, Niehaus, 2019).

Pritchett évoque l'importance de l'absence de théories communes sur le changement dans certains de ses écrits et discours : « Ce qui m'inquiète dans le développement, c'est qu'il y a deux catégories ontologiquement différentes auxquelles le mot est communément appliqué » (Pritchett, 2010a) ; « Vous pouvez appeler cela comme vous voulez, mais ne l'appelez pas développement » (Ogden, 2017 : 141). Mais, en général, l'absence de cet univers ontologique commun ne reçoit pas l'attention qu'elle mérite dans de tels débats. C'est peut-être parce que ces débats n'ont probablement que peu de valeur eux-mêmes – les intervenants utilisent les mêmes mots pour désigner des choses différentes.

L'évolution du « mouvement »

Parmi les plus éminents détracteurs du mouvement RCT figure Lant Pritchett. En 2018, Pritchett a commencé à donner une conférence qu'il a intitulée : « Le débat est terminé. J'ai gagné. Ils ont perdu »[5] (« *The Debate is Over. I won. They lost* »). Dans cette section, je ferai valoir que, après examen de l'évolution de

5. Sur la diapositive du titre lors de la conférence, on lit : « Nous avons gagné. Ils ont perdu », mais, dans ses remarques, il utilise le « je ».

l'utilisation et de la pratique des RCT, il apparaît clairement que de nombreuses critiques des RCT ont en fait été admises et validées en raison de l'évolution des pratiques parmi les praticiens des RCT et les centres de recherche. En ce sens, le titre triomphaliste de Pritchett est justifié. Toutefois, je soutiendrai également que l'évolution du « mouvement RCT » s'explique mieux par un modèle que Pritchett lui-même (ainsi que Matt Andrews et Michael Woolcock) a introduit (Andrews *et al.*, 2012) en arguant qu'il constituait la meilleure voie à suivre pour obtenir un impact durable sur le développement.

Adaptation itérative pour la résolution des problèmes

Dans leur article original – qui a donné lieu à un certain nombre d'autres articles ainsi qu'à un livre, *Building State Capacity* (Andrews *et al.*, 2017) –, Andrews *et al.* présentent les principes de l'adaptation itérative pour la résolution des problèmes (Andrews *et al.*, 2012 : *abstract* non paginé) :

> « Nous proposons une approche, l'adaptation itérative pour la résolution des problèmes (*Problem-Driven Iterative Adaptation* – PDIA), basée sur quatre principes fondamentaux, chacun d'entre eux contrastant vivement avec les approches standards. Premièrement, la PDIA se concentre sur la résolution de problèmes de performance localement identifiés et définis (par opposition à la transplantation de solutions préconçues de "bonnes pratiques"). Deuxièmement, elle cherche à instaurer un "environnement favorable" à la prise de décision qui encourage la "déviance positive" et l'expérimentation (par opposition à la conception de projets et de programmes que les agents sont obligés de mettre en œuvre en l'état). Troisièmement, elle inscrit cette expérimentation dans des boucles de rétroaction (*feedback loops*) étroites qui facilitent un apprentissage expérientiel rapide (par opposition aux longs délais d'apprentissage à partir d'une "évaluation" *ex post*). Quatrièmement, elle engage activement de larges groupes d'agents pour s'assurer que les réformes sont viables, légitimes, pertinentes et soutenables (par opposition à un groupe restreint d'experts externes qui encourage la diffusion "*top down*" de l'innovation). »

Ces quatre principes sont clairement visibles dans l'évolution du mouvement RCT.

Premier principe : résoudre des problèmes de performance localement identifiés et définis

On attribue généralement à Michael Kremer le mérite d'avoir commencé à utiliser les RCT dans le domaine du développement par une expérimentation randomisée sur l'impact des manuels scolaires sur l'apprentissage dans les écoles primaires kenyanes ; la seule controverse à cet égard est que Santiago Levy déployait presque simultanément une RCT pour évaluer l'impact de Progresa, un nouveau programme de transfert conditionnel de fonds au Mexique. Peu importe lequel des deux a réellement commencé l'utilisation, les deux projets incarnent

le premier principe. Levy avait pour ambition de résoudre un problème local particulier. Au moment où le programme a été créé, le gouvernement mexicain en place n'avait aucune chance de remporter les prochaines élections nationales. Levy craignait que le programme ne soit annulé par le prochain gouvernement. Il a lancé une RCT pour établir l'impact du programme et, ce faisant, le mettre à l'abri de toute interférence politique[6].

La première RCT de Kremer a été motivée par une discussion avec l'un de ses amis kenyans, Paul Lipeyah, à qui incombait la tâche de sélectionner sept écoles primaires pour recevoir de nouveaux manuels scolaires de l'ONG ICS. Kremer décrit ainsi son raisonnement :

> « En 1994, lorsque j'ai commencé à travailler au Kenya, j'ai été très influencé par le mouvement pour une meilleure identification [causale] dans l'économie du travail et les finances publiques […] Je ne réagissais pas non plus aux critiques des variables instrumentales. En effet, je pense que ceux qui travaillent sur des variables instrumentales et ceux d'entre nous qui travaillent sur des RCT ont été motivés par la même motivation, la crainte que beaucoup de travaux empiriques en économie à l'époque soient potentiellement la proie de variables de confusion et nécessitent pas mal d'hypothèses assez fortes. Cela étant, ce n'est pas comme si les variables instrumentales faisaient disparaître tous les problèmes, pas plus les RCT que d'autres. Je ne crois pas que quiconque pense que les RCT sont magiques, mais elles constituent un outil vraiment utile pour se faire une idée de l'impact causal. Je dirais donc que j'essayais de me faire une idée de l'impact causal d'une manière qui s'inscrivait dans un mouvement plus large de la profession économique afin de parvenir à une meilleure identification […] Ma principale motivation était d'ordre pratique – il s'agissait d'obtenir des réponses plus crédibles à des questions du monde réel. J'ai toujours été principalement intéressé par les questions sous-jacentes des politiques de lutte contre la pauvreté et j'ai compris que les RCT étaient un outil qui pouvait être adapté pour aider à répondre à cette question. J'avais l'ambition de faire des RCT un outil plus souple et plus utile » (Ogden, 2017 : 19).

De toute évidence, Kremer essayait lui aussi de résoudre un problème localement identifié et défini suivant deux axes. Premièrement, il essayait d'aider à résoudre le problème spécifiquement défini au niveau local par ICS, à savoir sélectionner les sept écoles qui devaient recevoir des manuels scolaires. Deuxièmement, il se penchait sur les problèmes localement identifiés et définis par les économistes pour améliorer l'identification causale.

Des histoires similaires circulent sur l'adoption des RCT par d'autres économistes, qui sont des utilisateurs bien connus de RCT. L'histoire de Pascaline Dupas, qui a ouvert ce chapitre, en est un bon exemple : la première RCT qu'elle

6. Entretien de l'auteur avec Levy en 2018.

a menée était en réponse au problème localement identifié et défini de savoir s'il valait mieux donner des moustiquaires ou les faire payer. La première RCT de David McKenzie, visant à tester les rendements du capital pour les micro-entrepreneurs au Sri Lanka, a vu le jour parce qu'il avait déjà fait un travail similaire non expérimental au Mexique, « mais les gens n'étaient pas convaincus par les résultats non expérimentaux » (Ogden, 2017 : 121). Il était motivé par le problème localement identifié et défini de convaincre des économistes et des décideurs politiques que les micro-entrepreneurs pouvaient arriver à un rendement du capital important. Parfois, comme dans l'histoire de Dupas et Kremer, le problème localement identifié et défini était une question posée à la fois par l'économiste et par une ONG ou une agence gouvernementale.

Deuxième principe : instaurer un « environnement favorable » à la prise de décision qui encourage la « déviance positive » et l'expérimentation

Il est clair que les premiers à recourir aux RCT ont instauré un environnement favorable qui a encouragé la déviance positive et l'expérimentation – au sens propre pour cette dernière. Le niveau d'innovation dans la conduite des RCT est assez impressionnant. À partir d'expériences généralement déployées à petite échelle sur des interventions très simples, par exemple la distribution de manuels scolaires dans sept écoles, les praticiens des RCT n'ont eu de cesse d'innover et d'évoluer. D'un point de vue méthodologique, le processus de randomisation et d'analyse est devenu beaucoup plus sophistiqué pour faire face à des contextes très variés et à des variables potentielles de confusion de l'équilibre prospectif, des tests multi-hypothèses et d'autres biais potentiels. Les premiers utilisateurs de RCT, tel Ted Miguel, ont été les pionniers de la transparence scientifique, de la mise à disposition des données et des analyses, et de l'utilisation de plans de pré-analyse. Une « deuxième vague » de *randomistas* a ensuite entrepris de mener des expérimentations beaucoup plus sophistiquées sur des périodes beaucoup plus longues (par exemple, les expériences de Blattman et Dercon (2018) comparant les emplois industriels au microcrédit) et à des échelles beaucoup plus vastes (par exemple, les expériences de Muralidharan et Niehaus avec National Rural Employment Guarantee Act [NREGA] et Aadhar) sur des sujets beaucoup plus complexes (par exemple, les expériences de Pomeranz sur les régimes fiscaux et les expériences de Karlan sur le contenu religieux d'une intervention).

Troisième principe : inscrire cette expérimentation dans des boucles de rétroaction étroites qui facilitent un apprentissage expérientiel rapide

Le seul argument concernant l'application de ce principe par les *randomistas* est de savoir s'il s'agit d'une chose qu'ils ont directement créée ou d'une caractéristique existante de l'enseignement économique qu'ils ont exploitée. Je dirais les deux. La nature de l'enseignement et de la pratique de l'économie est telle que chaque nouvelle génération d'économistes apprend en faisant le « sale boulot » des générations précédentes. Il est difficile d'imaginer une boucle de rétroaction plus étroite et un apprentissage expérientiel plus rapide qu'un doctorant (ou un étudiant à un stade moins avancé), qui passe un an ou deux en qualité

de responsable de recherche sur le terrain à travailler sur diverses expériences supervisées par des praticiens en place. Les utilisateurs des RCT ont également exploité le réseau d'événements dont dispose la profession économique comme une opportunité de boucles de rétroaction sur les innovations en matière de paramétrage, de gestion et d'analyse des expériences sur le terrain – des conférences telles que celles du Northeastern Universities Development Consortium (NEUDC) et du Pacific Development Consortium (PacDev) sont, au grand dam des détracteurs des RCT, souvent devenues des plateformes d'apprentissage rapide sur la manière de conduire, d'analyser et de rendre compte des RCT.

Quatrième principe : engager activement de larges groupes d'agents pour s'assurer que les réformes sont viables, légitimes, pertinentes et soutenables

Les institutions chargées de soutenir la mise en œuvre des RCT constituent, dans la pratique, les meilleurs exemples de ce principe. Depuis les premières RCT, les principaux utilisateurs de RCT ont créé des organisations comme les J-PAL, Innovations for Poverty Action (IPA) et Center for Effective Global Action (CEGA), qui se prêtent aisément à la description de larges groupes d'agents garantissant la viabilité, la légitimité, la pertinence et la soutenabilité des réformes. Toutes ces organisations sont impliquées dans des projets visant à atténuer les obstacles à la conduite et à la publication de RCT. Il s'agit notamment de livres et de cours sur la manière de mener des RCT, de la formation de nombreux étudiants au sein et hors des universités, et de la création d'organismes de recherche dans les pays en développement (par exemple, le Busara Center for Behavioral Economics et le personnel permanent déployé sur le terrain dans un certain nombre de pays).

Évolution et critiques des RCT fondées sur la PDIA

Conformément à la promesse d'Andrews, Pritchett et Woolcock, la mise en œuvre de la PDIA a permis d'améliorer considérablement la pratique des RCT, leur pertinence pour la pratique du développement et pour les décideurs politiques, mais aussi d'institutionnaliser le processus de conduite et de diffusion des résultats des expérimentations randomisées. Mais deux points supplémentaires méritent d'être mentionnés.

Premièrement, la pratique des RCT se développe au fur et à mesure que les praticiens abordent les problèmes qu'ils perçoivent. Les problèmes « localement identifiés » auxquels l'évolution est confrontée sont principalement les problèmes rencontrés par les praticiens des RCT – publication de recherches et réalisation d'objectifs de carrière personnels. Comme je le dirai plus tard, nombre de ces objectifs de carrière visent à améliorer le monde et à aider les plus pauvres. Mais cela n'empêche pas que le processus d'évolution soit animé par des motivations internes au mouvement.

Deuxièmement, ce processus d'amélioration interne est, selon Andrews, Pritchett et Woolcock, le seul moyen fiable et durable de renforcer les capacités. Pritchett

déplore souvent que nombre des problèmes localement identifiés que les *randomistas* ont tenté de résoudre soient entièrement prévisibles (Ogden, 2017) et que, malgré cela, les *randomistas* aient ignoré les critiques. Mais, comme le dit la citation ci-dessus sur l'introduction de la PDIA, « la transplantation de solutions préconçues de "bonnes pratiques" » et le changement basé sur un « groupe restreint d'experts externes qui encouragent la diffusion "*top down*" de l'innovation » ne fonctionnent tout simplement pas. La seule façon pour le mouvement RCT d'évoluer en une force de développement durable et efficace était de développer des capacités et des solutions en interne.

Dans cette section, j'examinerai brièvement comment le processus PDIA a conduit à l'évolution de la pratique des RCT pour répondre, au moins en partie, aux nombreuses critiques du mouvement.

Rien de magique

Comme indiqué dans une section précédente, nul ne sait exactement combien de praticiens des RCT ont jamais cru que les RCT étaient magiques ou ne faisaient l'objet d'aucun biais. Il convient de noter que Brodeur *et al.* (2018) trouvent beaucoup moins de preuves de *p-hacking*[7] et de recherche de significativité dans les articles mobilisant des RCT et ou des approches de discontinuité de la régression (RDD) que dans les articles ayant recours aux techniques de variable instrumentales ou de « différences de différence » et que Vivalt (2019) trouve que la significativité tend à être moins « gonflée » dans les RCT que dans les articles utilisant d'autres méthodes. Peut-être plus important encore pour la présente discussion, elle constate que cette tendance au « gonflement » de la significativité diminue au fil du temps.

Il est cependant probable que de nombreux praticiens des RCT n'aient pas apprécié, à leurs débuts, les nombreuses sources de biais qui subsistent dans les expérimentations randomisées. S'il y avait des gens pour croire que les RCT ont résolu tous les problèmes dénoncés par les critiques, on s'attendrait à ce que les praticiens des RCT résistent aux innovations de mise en œuvre et d'analyse qui prennent mieux en compte ces sources de biais potentiels.

De fait, ce que nous voyons, c'est que de nombreux économistes, que l'on pourrait qualifier de *randomistas*, innovent activement pour régler les problèmes de biais, de fiabilité et de réplicabilité. Plusieurs d'entre eux méritent une mention particulière.

> Ted Miguel est l'un des fondateurs de l'Open Science Framework et de la Berkeley Initiative for Transparency in Social Sciences. Ces deux organisations encouragent les chercheurs à rendre accessibles toutes les données et tous les codes utilisés dans leurs travaux afin d'autoriser leur réplication (qu'il s'agisse ou non de RCT).

7. N.D.E. : manipulation inappropriée de l'analyse des données pour permettre à un résultat favorable d'être présenté comme statistiquement significatif.

Certains *randomistas*, dont Dean Karlan, Esther Duflo et Chris Blattman, se sont fait les avocats du pré-enregistrement des études, en particulier des RCT, et ont participé à la création du registre RCT de l'American Economic Association (AEA).

Le J-PAL a mis en place un service de réplication où « un étudiant diplômé tente de répliquer un papier complet à partir de zéro et peut identifier toute erreur, omission ou hypothèse douteuse » (Crépon *et al.*, 2019 : 2).

Le Development Impact Blog de la Banque mondiale (« *news, methods and insights about impact evaluation* ») poste fréquemment des critiques d'articles évoquant des RCT, ainsi que des conseils sur les nouvelles méthodes et techniques statistiques pour améliorer l'analyse des RCT et sur des méthodes non RCT.

Guido Imbens, cité plus haut comme ayant déclaré que les RCT siègent en fait au sommet de la hiérarchie des méthodes, continue à travailler sur les améliorations méthodologiques d'autres méthodes, comme la méthode des doubles différences et l'apprentissage automatique (par exemple, Athey et Imbens, 2018 ; 2019).

Chacun de ces cas peut être considéré comme un exemple de chaque principe de la PDIA, étant donné qu'ils abordent tous des problèmes de performance dans la conduite et l'interprétation des RCT et sont menés sans aucun pouvoir ou mandat d'autorisation centralisée. Plus important encore, ces efforts montrent que, loin de traiter les RCT comme de la magie, ceux qui se positionnent dans le giron aux contours flous du mouvement RCT reconnaissent des sources de biais et d'erreurs dans les RCT et investissent activement pour y remédier. De même, ils continuent d'employer des méthodes autres que les RCT et n'ignorent pas, en pratique, les travaux sur d'autres méthodes.

Boîte noire

Une fois de plus, l'application des RCT suivant les principes de la PDIA pour tenir compte des limites de RCT simples dans l'exposition des mécanismes causaux a considérablement évolué. En voici quelques exemples :

Alfonsi *et al.* (2017) étudient des programmes d'emploi des jeunes en Ouganda en comparant des programmes de formation professionnelle aux subventions pour la formation en cours d'emploi. Ils établissent non seulement que la formation professionnelle a un impact plus important en termes de revenus des jeunes, mais aussi que le mécanisme autorise une plus grande mobilité entre employeurs grâce à des compétences certifiables, ce qui conduit au final à une meilleure adéquation entre les jeunes diplômés et les entreprises à plus forte productivité.

Beaman *et al.* (2018a) étudient la diffusion des technologies agricoles *via* les réseaux sociaux au Malawi et établissent la nature de la « contagion complexe » qui conduit les agriculteurs à adopter de

nouvelles méthodes. Ils utilisent ce mécanisme pour identifier des moyens d'améliorer de manière rentable le ciblage des programmes de vulgarisation agricole.

Cai et Szeidl (2018) mènent une expérimentation avec des réseaux d'entreprises chinois et constatent que les réunions de réseautage améliorent considérablement les performances des entreprises, notamment grâce à l'apprentissage par les pairs dispensé par les entreprises les plus performantes sur des sujets indépendants de l'intervention et à une meilleure adéquation entre les fournisseurs et les clients ; ils notent aussi que la régularité des réunions est importante.

Campos *et al.* (2017) étudient des programmes de formation destinés aux petites entreprises du Togo en comparant la formation professionnelle traditionnelle à l'instruction sur l'initiative personnelle. Ils constatent non seulement que la formation sur l'initiative personnelle est plus efficace pour accroître les bénéfices de l'entreprise, mais aussi que les comportements spécifiques ont changé notamment la diversification des produits, l'innovation et l'investissement.

Karing (2018) étudie un programme visant à encourager la vaccination des enfants en Sierra Leone et établit non seulement l'efficacité de la signalisation publique par le biais de bracelets colorés, mais constate aussi que le mécanisme est la désirabilité sociale (et non l'attention) et que l'effet des bracelets varie en fonction de la désirabilité sociale de vaccins spécifiques.

Les exemples choisis ici ne sont pas systématiques, mais visent à démontrer que les efforts déployés pour répondre à la critique « boîte noire » ne sont pas l'apanage de quelques chercheurs, écoles, contextes ou secteurs. Au fur et à mesure que des études de ce type sont conduites, le processus PDIA opérant implicitement au sein du mouvement RCT signifie que les nouvelles RCT seront de plus en plus amenées à établir des mécanismes de causalité.

Validité externe

La critique de la validité externe serait globalement plus crédible si certains de ses partisans s'exprimaient aussi clairement sur les problèmes de validité externe de toutes les études, et non pas seulement des RCT. Comme l'a fait remarquer Pam Jakiela en réponse à Cartwright et Deaton, « joli aperçu de Deaton et Cartwright mais, pour une raison quelconque, le terme "étude" s'écrit pour eux "R-C-T." »[8]. Cela étant, de nombreux promoteurs de RCT prodiguent des conseils politiques, ce qui présuppose une validité externe.

Face au problème de la preuve de la validité externe, les praticiens des RCT ont évolué diversement. Ils se sont d'abord employés à déterminer empiriquement si les résultats des RCT dans un contexte donné pouvaient prédire les résultats

8. https://twitter.com/PJakiela/status/797053999925104640

dans un autre. Par exemple, MEAGER (2019), avec la modélisation hiérarchique bayésienne (un autre exemple de l'application de la PDIA à la critique « rien de magique »), montre que la variation entre des RCT sur le microcrédit est plus ténue qu'il n'y paraît (PRITCHETT et SANDEFUR, 2015) et donc que les résultats des RCT individuelles sont raisonnablement prédictifs des résultats en d'autres lieux. ALCOTT (2015) fait quelque chose de similaire en comparant la capacité d'une RCT portant sur les rappels à réduire la consommation d'énergie dans une ville à prédire l'effet de la même campagne dans une autre ville ; il a constaté que les RCT ne sont pas très efficaces, mais le sont plus que d'autres méthodes d'usage courant.

Dans le même temps, les praticiens des RCT se sont beaucoup plus intéressés aux réplications et aux études pluralistes dans de multiples contextes. Par exemple, DUPAS *et al.* (2018) testent les incitations à l'épargne en Ouganda, au Malawi et au Chili. Plus marquant encore, des chercheurs issus de différentes disciplines ont collaboré pour tester des programmes « *Targeting the Ultra Poor* » dans huit endroits distincts, avec des mises en œuvre gouvernementales et d'ONG. Cette « réplication » à des fins de validité externe peut également se faire *ex post*. Par exemple, BERNHARDT *et al.* (2017) réanalysent les données de multiples expériences sur les différences de rendement du capital entre des hommes et des femmes entrepreneurs afin d'identifier un mécanisme causal, qui était jusqu'alors resté flou, concernant la négociation et l'optimisation des ménages (ce qui est également un exemple de réponse à la critique de la boîte noire). Ces réplications *ex post* sont facilitées par les efforts déployés par d'autres utilisateurs de RCT pour garantir la disponibilité des données et des codes de toutes les expérimentations à des fins de réplication.

Bien entendu, l'application de toute évaluation d'impact (RCT ou autre) pour prédire les résultats dans d'autres contextes posera toujours des questions de validité externe. Mais des approches plus systématiques sont aussi en pleine évolution. Plus les RCT s'intéresseront aux mécanismes de causalité, plus les hypothèses sur la validité externe deviendront explicites, et plus d'études comporteront des modèles structurels. Des cadres plus formels pourront ainsi être conçus afin d'évaluer la validité externe et d'intégrer les résultats de multiples études telles que DEHEJIA *et al.*, (2019) et WILKE et HUMPHREYS (2019).

Signification négligeable

J'ai indiqué plus haut qu'un des principaux fondements de la critique « signification négligeable » est la différence de théories du changement entre les *randomistas* et les critiques. Il est difficile de répondre quoi que ce soit à la critique selon laquelle les seuls changements qui comptent sont les politiques au niveau macro. Cette critique porte toutefois plus sur la micro-économie appliquée en général que sur les RCT. Cela étant, le mouvement RCT a une réponse à au moins une des critiques émanant d'une théorie du changement différente. Pour ceux qui sont d'avis que les institutions comptent, les prouesses réalisées par les *randomistas* pour renforcer les institutions ont de quoi impressionner. Outre les exemples évidents de l'IPA et du J-PAL, les institutions bâties par le mouvement

randomista incluent directement et indirectement le Global Innovation Fund, le Busara Center for Behavioral Research, l'International Initiative for Impact Evaluation (3ie), Evidence Action, Development Impact Ventures, *American Economic Journal : Applied Economics* (*AEJ:AE*) et de nombreux cabinets d'études locaux.

Je souhaite cependant me concentrer ici sur diverses critiques « importance négligeable », qui se fondent sur les expériences originales ayant popularisé l'utilisation des RCT – manuels scolaires, formation des enseignants, incitation à la vaccination, etc. Ces critiques portent à la fois sur la nature restreinte de l'intervention et sur les petits résultats mesurés, même lorsqu'ils sont statistiquement significatifs (voir l'exemple de Harrison, 2011). Une autre variation connexe déplore que les RCT ne soient pas très bien adaptées pour mesurer l'impact à long terme (Ravallion, 2020).

Ces problèmes ont suscité une impressionnante créativité dans l'application des RCT. La réponse la plus directe à la critique du « trop petit » a consisté à élargir l'échelle des RCT. Muralidharan *et al.* (2016 ; 2018c) offrent le meilleur exemple à cet égard en étudiant un programme de filet de sécurité dans l'Andhra Pradesh, en Inde – expérience regroupant 19 millions de personnes –, pour des économies estimées à 38,5 millions de dollars par an. Le fait que ces deux chiffres soient exprimés en millions plutôt qu'en milliards en laissera certains sur leur faim, mais c'est assurément un progrès notable observé au cours des premières années du mouvement RCT.

D'autres *randomistas* ont repoussé les limites de ce qui peut être étudié par les RCT d'autres manières :

– Dina Pomeranz, avec plusieurs co-auteurs, a conduit des RCT sur diverses questions de politique fiscale ;

– Chris Blattman, avec plusieurs co-auteurs, a randomisé l'accès aux emplois d'usine en Éthiopie, des stratégies de police en Colombie et des campagnes de lutte anti-violence au Liberia ;

– Gharad Bryan, James Choi et Dean Karlan vont même jusqu'à réaliser une RCT sur l'impact de la croyance religieuse, en randomisant l'évangélisation chrétienne aux Philippines (étayant là l'hypothèse protestante de l'éthique du travail).

Saucisse politique

La traduction des résultats des RCT en changements politiques a toujours été un objectif explicite des praticiens des RCT. Quand ceux-ci doivent expliquer comment et pourquoi ils se sont lancés dans les RCT, ils expriment tous le désir pragmatique d'influencer la politique afin de faire une différence concrète dans la vie des gens[9]. Une citation attribuée à Michael Kremer par Karthik Muralidharan

9. Dans Ogden (2017), voir les interviews de Michael Kremer, Esther Duflo, Dean Yang, Chris Blattman, entre autres.

illustre bien ce point : « Ne vous excusez jamais quand vous dites que votre motivation première est d'améliorer la vie de centaines de millions de personnes et que l'économie est un outil pour y parvenir, et non une fin en soi[10]. »

En dépit de leurs bonnes intentions, le travail initial des *randomistas* en termes d'influence politique pourrait être décrit comme l'a fait Pritchett en diverses occasions : naïf, préhistorique. Comme le notent BÉDÉCARRATS *et al.* (2019b), moins de 5 % des évaluations d'impact des RCT pratiquées par le J-PAL ont conduit à des changements politiques à grande échelle.

Mais, au fil du temps, la sophistication et l'intensité des efforts déployés pour influer sur les politiques se sont accélérées. Après avoir cerné ce problème localement identifié à impact politique limité, les praticiens des RCT ont rapidement répété l'expérience dans un environnement propice à la déviance positive et ont vite dégagé des leçons utiles pour l'ensemble de la communauté des *randomistas*.

L'hypothèse initiale, à savoir que les preuves empiriques généreraient mécaniquement un changement politique, a été abandonnée au profit d'efforts ciblés visant à influencer la politique. Des équipes axées sur les politiques ont ainsi été mises sur pied au sein du J-PAL (y compris une initiative « innovation gouvernementale », qui s'emploie spécifiquement à soutenir les agences gouvernementales menant des expériences de mise en œuvre de politiques) et d'IPA. Mais ces efforts supposent également une participation à la création d'organisations autonomes pour mettre en œuvre des programmes basés sur des preuves empiriques issues de RCT (en l'occurrence, Evidence Action), d'organisations visant à encourager la création et l'utilisation de preuves empiriques dans l'élaboration des politiques (3ie), de groupes internes au sein d'organisations existantes chargées de l'élaboration et de la mise en œuvre des politiques (Development Impact Ventures de l'United States Agency for International Development [USAID]), d'une collaboration étroite avec des groupes de recherche d'ONG (Building Resources Across Communities [BRAC], Pratham), de programmes d'éducation destinés aux décideurs et responsables de la mise en œuvre des politiques et, bien sûr, de la formation d'un grand nombre d'étudiants de master et de doctorat aux méthodes et approches, de façon à ce que la grande majorité d'entre eux trouvent des emplois dans le monde politique plutôt que dans le milieu universitaire. Certains praticiens ont même assumé des fonctions au sein de l'appareil politique – je pense notamment à Rachel Glennerster, économiste en chef au Department for International Development (DFID), et aux fonctions parlementaires d'Andrew Leigh en Australie.

Autrement dit, les *randomistas* ont engagé un large éventail d'agents pour garantir la validité et la continuité de l'utilisation des RCT en vue d'influencer les politiques. Une nouvelle génération de praticiens des RCT va être amenée à faire partie intégrante des institutions chargées de l'élaboration des politiques (ne serait-ce qu'en raison de la pénurie d'emplois dans les universités).

10. https://twitter.com/karthik_econ/status/1102237584103600129

Les critiques éthiques

Il y a beaucoup moins à dire sur ce sujet. C'est en partie dû au fait que, fondamentalement, les *randomistas* sont convaincus que l'expérimentation sur les êtres humains est éthique, indépendamment des intuitions morales de la majorité du public américain, une attitude bien sûr partagée par la plupart des scientifiques. Le refrain commun, que je partage aussi, est qu'il n'y a pas de choix quant à l'opportunité d'expérimenter (puisque toute mise en œuvre politique est une expérimentation) – il y a seulement un choix quant à la quantité d'enseignements à tirer d'une expérimentation.

Cela étant, il subsiste naturellement de nombreuses questions sur l'éthique de l'expérimentation. Durant l'été 2019, un nouveau document de travail qui a randomisé l'incitation à participer aux manifestations anti-autoritaires à Hong Kong (Bursztyn *et al.*, 2019) a attiré une grande attention[11], notamment parce que de nombreux économistes semblent trouver l'expérience contraire à l'éthique. Une question souvent posée, du moins sur Twitter, visait à déterminer comment l'expérience avait apparemment réussi à passer entre les mailles du filet de plusieurs commissions éthiques. Le document et la discussion qui a suivi ont révélé[12] qu'il n'existe pas encore de limites ou de codes significatifs, ni même de principes partagés sur les limites éthiques que les économistes devraient fixer en matière d'expérimentation.

En ce qui concerne les questions d'équipoise, comme indiqué plus haut, il s'agit d'un domaine dans lequel le mouvement RCT ne s'est pas encore significativement engagé, pour autant que je sache.

Trop : la dernière critique

La dernière catégorie de critiques que j'ai identifiée sort du cadre de la PDIA, car elle ne critique pas les travaux des praticiens des RCT en économie du développement, mais le volume et le nombre de ces travaux. Je trouve que c'est la moins convaincante de toutes les critiques dans le domaine économique.

Pour commencer, comme le reconnaissent de nombreuses critiques « trop », l'émergence des RCT dans l'économie du développement est due en grande partie aux conditions et à la structure du marché de l'économie académique. L'utilisation des RCT a gagné en popularité dans un contexte de questions fréquentes sur la crédibilité d'autres méthodes, dans un environnement qui exigeait des aspirants économistes qu'ils livrent un travail crédible, inédit et publiable. Les RCT ont promis – et livré – un travail réunissant ces trois qualités. Ainsi, la critique « trop » devrait vraiment s'adresser aux structures et aux incitations de la profession, et non à ceux qui répondent aux incitations générées par la profession. Cette forme de critique équivaut à critiquer les acteurs du marché pour avoir fait la « mauvaise » chose, plutôt que de s'attaquer aux éventuelles défaillances du marché.

11. https://twitter.com/DurRobert/status/1148090885470654464
12. https://twitter.com/arindube/status/1148807790787473410

Deuxièmement, la critique « trop » ne parvient pas à articuler une mesure objective de ce que pourraient être les seuils entre « pas assez », « juste ce qu'il faut » et « trop ». Il est objectivement vrai que l'utilisation des RCT et la publication d'articles utilisant la méthode ont fortement augmenté (Ravallion, chap. 1, ce volume ; Bédécarrats *et al.*, 2019b : 489), mais cette croissance doit être mise en perspective. Il convient ici de citer l'étude de McKenzie (2019) qui examine en détail les données sur cette question :

> « […] malgré la croissance rapide, la majorité des articles sur l'économie du développement publiés dans les cinq plus grandes revues ne sont pas des RCT […] Sur les 454 articles de développement publiés dans ces 14 revues [consacrées au développement économique] en 2015, seuls 44 sont des RCT (9,7 %). La conséquence est que les études RCT ne représentent qu'une petite partie de l'ensemble de la recherche en cours sur le développement.
>
> Le chercheur médian [affilié au Bureau for Research and Economic Analysis of Development (BREAD)] avait publié neuf articles, et la part médiane de leurs articles parlant de RCT était de 13 %. En se concentrant sur le sous-ensemble de ceux qui ont publié au moins une RCT, le pourcentage moyen (médian) de leurs papiers RCT publiés est de 35 % (30 %) et la fourchette de 10-90 est de 11 à 60 %. Ainsi, les jeunes chercheurs qui publient des RCT écrivent et publient également des articles qui ne parlent pas de RCT. »

Troisièmement, la prémonition souvent répétée selon laquelle « l'enthousiasme pour les RCT va s'estomper » me semble être une critique creuse. Bien sûr, nous devons nous attendre à ce que les méthodes continuent de s'améliorer, à ce que des innovations de toute sorte mettent au jour des problèmes jusqu'ici méconnus et améliorent les approches. Dans un avenir pas trop lointain, je peux prédire avec confiance que quelqu'un écrira un essai sur les « RCT 2.0 » et établira une distinction discutable entre les « premiers jours » du mouvement RCT et les méthodes améliorées alors en vogue. Ce chapitre entre peut-être dans cette catégorie.

Susan Athey, en réponse à Judea Pearl critiquant ce qu'il appelle l'approche naïve de l'inférence causale en économie (qui, dans son ensemble, ne relève pas du mouvement RCT), écrit ceci :

> « [Je] pense que la manière la plus efficace d'évangéliser une nouvelle méthode est de démontrer son efficacité dans une application empirique de premier ordre, où la méthode aboutit clairement à un résultat de meilleure qualité et plus crédible. Les chercheurs imiteront un exemple de réussite pleinement élaboré[13]. »

13. https://twitter.com/Susan_Athey/status/1107422021753790464

Cette citation dresse à elle seule l'historique de l'utilisation des RCT dans l'économie du développement. L'enthousiasme pour la pratique originale des RCT s'est déjà estompé avec l'apparition des « applications empiriques de premier ordre » d'expériences et d'analyses plus sophistiquées. Et l'enthousiasme pour la pratique actuelle s'estompera certainement au fur et à mesure que naîtront des « applications empiriques de premier ordre » de méthodes améliorées – avec ou sans randomisation. D'ici là, il n'y a pas de « trop » qui vaille.

Enfin, certains déplorent que les « meilleurs et les plus brillants » économistes gaspillent leurs talents en se concentrant exclusivement sur les RCT. Cette critique est la moins logique de toutes. Si les détracteurs ont raison, que les problèmes des RCT sont insurmontables, et qu'il existe clairement de meilleures alternatives, cela indique nécessairement que ceux qui continuent à utiliser principalement des RCT ne sont pas les plus brillants et les meilleurs. Cette critique doit expliquer pourquoi il faudrait croire que les plus brillants et les meilleurs se trompent systématiquement et demeurent néanmoins dignes de leur réputation. Et s'ils ne sont pas les plus brillants et les meilleurs, pourquoi les économistes qui sont véritablement les plus brillants et les meilleurs ne peuvent-ils pas convaincre la prochaine génération d'étudiants de renoncer aux RCT au profit d'autres méthodes ? Seule explication plausible de cette critique, c'est toute la profession économique qui est défaillante, auquel cas les détracteurs perdent leur temps à s'intéresser aux symptômes et non aux causes.

Conclusion

Pour conclure, je souhaite établir un cadre de réflexion distinct sur l'évolution de la pratique des RCT et sur les différentes critiques et réponses. Ici, une fois de plus, je rejoins l'analyse de Pritchett. J'ai passé les dix premières années de ma carrière au sein de la société de recherche technologique Gartner. L'un des produits les plus connus de cette organisation est le « cycle du *hype* » – une façon de concevoir l'émergence, l'évolution et l'adoption des technologies émergentes.

Le cycle du *hype* postule que, lorsqu'apparaît une technologie révolutionnaire, elle passe par cinq phases distinctes, nommées de façon suffisamment colorée pour être suffisamment explicites : « déclencheur d'innovation » (*Innovation Trigger*), « pic des attentes exagérées » (*Peak of Inflated Expectations*), « gouffre de la désillusion » (*Trough of Disillusionment*), « pente de l'illumination » (*Slope of Enlightenment*) et « plateau de la productivité » (*Plateau of Productivity*) (fig. 1).

Pritchett est tombé par hasard sur le cycle du *hype* et l'a appliqué aux RCT dans un essai de 2013. Je reconnais qu'il s'agit d'un modèle utile pour réfléchir aux RCT – de fait, je dirais qu'il vaut mieux considérer les RCT comme une « technologie émergente » dans l'économie du développement, plutôt que comme un mouvement.

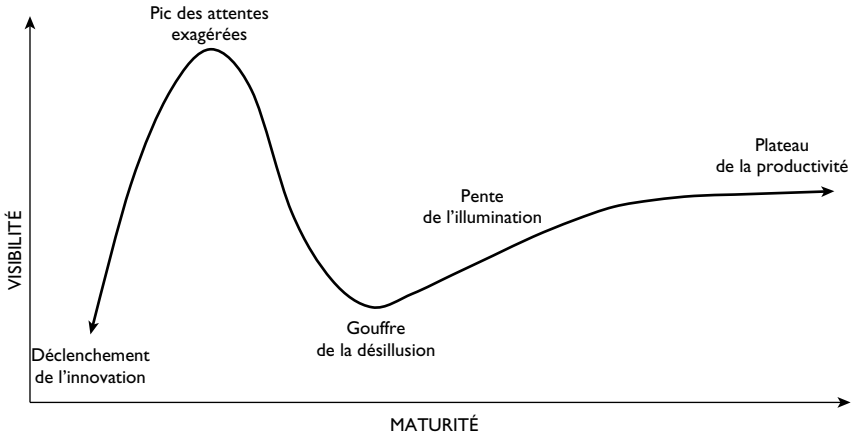


Figure 1
Le cycle de hype de Gartner.
Source : Timothy Ogden, d'après le cycle de *hype* de Garner.

Sous le prisme du cycle du *hype*, j'ai démontré dans ce chapitre que (1) le pic des attentes exagérées concernant les RCT était certes réel, mais qu'il n'a jamais été aussi élevé que les critiques l'avaient laissé entendre et que, en tout état de cause, il est aujourd'hui révolu ; (2) l'enthousiasme initial pour les RCT a rapidement été satisfait par une série de critiques valables, conduisant, sinon à un gouffre de la désillusion, du moins à des changements significatifs dans l'utilisation et la pratique des RCT ; et (3) la situation présente est clairement sur la pente de l'illumination, comme en témoignent les déclarations et pratiques des utilisateurs avancés de la technologie.

Il convient de noter spécifiquement que l'évolution de la pratique des RCT valide bon nombre des critiques détaillées ici. L'évolution que j'ai tenté de documenter répond à ces critiques. Les praticiens des RCT ne font pas évoluer leur pratique pour traiter de questions nouvelles qui n'ont pas été soulevées par les critiques – indépendamment de toute réponse directe à leurs détracteurs, les *randomistas* ont implicitement admis bon nombre des critiques en évoluant de telle façon que beaucoup d'entre elles ont perdu de leur pertinence.

Cela étant, je pense qu'il y a de bonnes raisons de croire que le plateau de la productivité des RCT est plus élevé que ce que de nombreux critiques semblent penser, simplement parce que c'est là une conséquence mécanique du fonctionnement du monde. Il y a beaucoup plus de décisions sur la mise en œuvre en soi que sur l'objet de la mise en œuvre. Les décisions de mise en œuvre relèvent clairement du champ d'application des RCT. Comme il y a beaucoup plus d'étudiants formés à l'économie du développement que d'étudiants qui occuperont un emploi permanent dans des universités doctorales, une proportion importante de ces étudiants finira par trouver un travail privilégiant les questions de mise en œuvre sur les questions de politique générale. La formation qui leur sera dispensée en matière d'expérimentation

et d'identification causale sera très spécialisée et leur permettra d'appliquer la PDIA dans leur travail.

En outre, les RCT sont un outil plus utile pour améliorer le monde que la plupart des outils à la disposition de l'*économiste du développement médian*, étant donné la nature et les exigences de la profession, ainsi que les difficultés de l'influence politique. L'émergence de la technologie RCT et les mécanismes qui étayent cette technologie sont applicables à la grande majorité des questions réelles et des décisions discrètes concernant les politiques, les programmes et la mise en œuvre de la lutte contre la pauvreté. Il est vrai que les RCT n'aident probablement pas à évaluer les politiques de taux de change, le niveau optimal de la dette publique ou les conséquences de l'inégalité des richesses (pour ne citer que quelques exemples). Mais les politiques liées aux réponses aux questions sur ces sujets se prêtent beaucoup moins à l'influence académique, quelle que soit la méthodologie utilisée pour y répondre. Au moment où j'écris ces lignes (à l'été 2019), la possibilité d'un repli massif des régimes commerciaux libéralisés est effroyablement réelle, en dépit des efforts politiques déployés par des dizaines de milliers de macroéconomistes depuis des décennies. Il n'y a aucune raison de croire que l'impact marginal de l'économiste du développement moyen qui étudie l'un de ces sujets est supérieur à un zéro estimé avec précision. L'avantage comparatif de l'économiste du développement médian résiderait dans l'amélioration de la mise en œuvre d'une politique ou d'un programme et ce, même sans validité externe ni transposition à plus grande échelle.

En conclusion, je ne peux que répéter une fois de plus que Lant Pritchett avait raison, et qu'il a gagné. Les critiques du mouvement RCT sont généralement valables, si ce n'est objectivement correctes. Toutefois, l'évolution de la pratique des RCT est telle qu'elle a répondu à bon nombre de ces critiques. Je m'attends à ce que les RCT poursuivent leur évolution et soient finalement supplantées par une autre méthodologie. Il existe déjà, bien sûr, de nouvelles « technologies émergentes » en économie : *big data*, apprentissage automatique et intelligence artificielle – et certains des débats sur l'utilisation et les applications des RCT sont en train de se répéter. D'ici là, je pense que le plateau de la productivité où se trouvent actuellement les RCT continuera d'apporter des bénéfices au monde entier.

Annexe au chapitre 4 : citations complètes

1. « Les meilleures méthodes à utiliser et leurs combinaisons possibles dépendent de la question exacte en jeu, du type d'hypothèses de base qui peuvent être raisonnablement employées et du coût des différents types d'erreurs » (Deaton et Cartwright, 2018 : 3).

2. « Nous ne devrions pas encourager ou décourager l'emploi d'un outil particulier juste pour ce qu'il est. Nous devrions encourager les étudiants à poser une question intéressante et à employer le bon outil pour y répondre. Point » (OGDEN, 2017 : 86).

3. « De très bonnes données descriptives qui attirent l'attention des gens sur quelque chose qui ne les avait jusque-là pas intéressés ont changé l'opinion des gens en matière de politique, autant que n'importe quelle expérience » (OGDEN, 2017 : 134).

4. « La nouveauté a peut-être conduit à une surenchère sur les RCT, comme ces déclarations stupides affirmant que tout doit être randomisé pour être évalué, ou les gens qui disent ne croire à aucune preuve issue de données observationnelles » (OGDEN, 2017 : 238).

5. « Je pense que ceux qui travaillent sur les variables instrumentales et ceux d'entre nous qui travaillent sur des RCT ont été motivés par la même motivation, la crainte que beaucoup de travaux empiriques en économie à l'époque soient potentiellement la proie de variables de confusion et nécessitent pas mal d'hypothèses assez fortes. Cela étant, ce n'est pas comme si les variables instrumentales faisaient disparaître tous les problèmes, pas plus les RCT que d'autres. Je ne pense pas qu'il faille croire que les RCT sont magiques » (OGDEN, 2017 : 19).

6. « […] Si je devais choisir, je dirais que nous devrions insuffler plus d'énergie dans les grandes choses que dans les petites. Je ne crois pas que ce soit en partie ainsi que le mouvement d'évaluation randomisée a été vendu aux décideurs politiques : "Vous allez avoir des réponses". Je ne crois pas que ce soit là ce que nous allons avoir. J'ai le sentiment que nous allons voir des évaluations partout dans le monde » (OGDEN, 2017 : 227).

7. « Les organisations devraient pouvoir puiser dans différents domaines pour répondre aux questions pertinentes… Je constate de nombreux recoupements entre différentes formes d'identification causale. Donc je pense qu'il faut se concentrer, certes, mais je ne pense pas qu'il faille se concentrer uniquement sur des évaluations randomisées. Je ne pense pas que cela fasse sens » (OGDEN, 2017 : 200).

8. « Une évaluation d'impact devrait permettre de déterminer pourquoi quelque chose fonctionne, et non pas simplement *si* cela fonctionne. Il ne faudrait pas faire d'évaluations d'impact si elles ne peuvent fournir aucune connaissance généralisable sur la question du "pourquoi", c'est-à-dire si elles ne sont utiles qu'à l'organisation chargée de la mise en œuvre et uniquement pour cette mise en œuvre donnée. Cette règle s'applique aux programmes ayant peu de possibilités de mise à l'échelle, peut-être parce que les bénéficiaires d'un programme particulier sont hautement spécialisés ou inhabituels, ou parce que le programme est rare et a peu de chances d'être répliqué ou étendu. Si les évaluations n'ont qu'une seule utilisation possible, elles n'en valent presque jamais le coût » (GUGERTY et KARLAN, 2018 : 45).

9. « Il y a de nombreux exemples banals et inutiles d'études appliquant chaque méthode spécifique » (McKenzie, 2018).

Remerciements

Merci à Jonathan Morduch, David McKenzie, Lant Pritchett et Isabelle Guérin pour les discussions utiles que nous avons eues sur l'élaboration de ce chapitre, ainsi qu'aux participants de l'atelier de l'AFD « *Randomized Control Trials in the Field of Development*: *The Gold Standard Revisited* » (Paris, mars 2019). Merci également à Cynthia Kinnan, Jessica Goldberg, Johannes Haushaufer, Cyrus Samii et Bruce Wydick pour leurs précieux conseils et aux trois éditeurs du livre, Florent Bédécarrats, Isabelle Guérin et François Roubaud, pour leurs précieux commentaires et discussions.

Partie 2

Perspectives sectorielles

Chapitre 5

Réduire le déficit des connaissances dans la prestation de soins de santé à l'échelle mondiale

Apports et limites des expérimentations aléatoires

Andres Garchitorena, Megan B. Murray, Bethany Hedt-Gauthier, Paul E. Farmer et Matthew H. Bonds

Contexte : les expérimentations aléatoires dans le domaine de la médecine et de la santé mondiale

La recherche d'une évaluation empirique, systématique et rigoureuse de l'efficacité des interventions destinées à la population humaine n'a jamais cessé d'être une préoccupation pour la médecine, bien avant toute autre discipline. L'évaluation par assignation aléatoire (*Randomized Controlled Trials* – RCT) est parmi ces démarches d'évaluation rigoureuses, dans laquelle les chercheurs affectent des statuts de traitement à des individus sélectionnés de manière aléatoire avant de comparer les résultats. Soutenus par une industrie pharmaceutique florissante, les essais contrôlés ont été progressivement adoptés au cours des XVIII^e^ et XIX^e^ siècles (Bothwell et Podolsky, 2016). Ceux-ci visaient à distinguer les produits médicaux efficaces (comme les vaccins, les antibiotiques) des nombreux remèdes, thérapies ou répliques aux effets positifs douteux (Bothwell et Podolsky, 2016). Au cours de la première moitié du XX^e^ siècle, les chercheurs

ont souvent eu recours à d'autres modèles d'affectation (en traitant par exemple un patient sur deux), mais ceux-ci entraînaient d'importants biais de sélection, les médecins sélectionnant les patients en fonction des besoins qu'ils percevaient. L'épidémiologiste Austin Bradford Hill s'est attaqué à cette problématique en 1948 en commençant une série d'essais sur le traitement de la tuberculose qui mettait en œuvre une randomisation aveugle stricte des patients (autrement dit, des « évaluations randomisées »). Soutenues par le British Medical Research Council et rapidement adoptées par la communauté des chercheurs, les RCT se sont vite hissées au premier rang des modèles expérimentaux en recherche clinique. En 1970, la Food and Drug Administration américaine a exigé de l'industrie pharmaceutique qu'elle fournisse des résultats de RCT avant d'autoriser la mise sur le marché de tout nouveau médicament, conférant alors un rôle central aux RCT dans les réglementations et directives internationales (Bothwell et Podolsky, 2016).

L'idée de promouvoir les RCT au rang d'étalon-or de la recherche clinique a été favorisée par un mouvement en faveur de la médecine basée sur les preuves (en anglais *evidence-based medicine* – EBM), qui vise à améliorer la pratique clinique par une évaluation critique de la littérature scientifique pour permettre aux cliniciens d'adopter les meilleures pratiques. Largement influencé par le livre d'Archie Cochrane intitulé *Effectiveness and Efficiency: Random Reflections on Health Services*, paru en 1972, la médecine basée sur les preuves s'appuie sur le classement hiérarchique de la qualité des études d'efficacité en fonction de la méthodologie utilisée, les RCT figurant en tête de ce classement et les études observationnelles sans groupe de contrôle en queue de peloton. Au cours des décennies suivantes, les RCT ont gagné en popularité hors du cadre de la recherche clinique. Dans les pays occidentaux, leur utilisation s'est étendue à l'évaluation des politiques publiques dans les domaines de l'éducation, de l'économie, de la sociologie et de la santé publique. Le développement des essais randomisés par grappes à la fin des années 1970, qui randomisaient des groupes de sujets plutôt que des individus isolés, a permis une application encore plus large aux évaluations pour lesquelles la randomisation individuelle était irréalisable ou indésirable. L'utilisation des RCT dans les domaines de la santé mondiale et du développement international a cependant pris du retard, avec seulement quelques dizaines d'études publiées avant les années 2000 (Cameron *et al.*, 2016).

Un tournant majeur a été pris au début du XXIe siècle avec l'établissement des Objectifs du millénaire pour le développement (OMD). Reconnaissant que la santé est à la fois un objectif central du développement et un moteur potentiel (Sachs, 2001), les Nations unies ont veillé à ce que les résultats en matière de santé y figurent en bonne place, avec des engagements à réduire la mortalité infantile (OMD 4), à améliorer la santé maternelle (OMD 5) et à combattre le sida, le paludisme et d'autres maladies (OMD 6). Les financements alloués à ces domaines sont montés en flèche (Institute for Health Metrics and Evaluation [IHME], 2016), les fondations privées ont commencé à jouer un rôle de plus en plus important et de grandes organisations comme GAVI, l'Alliance du vaccin, et le Fonds mondial de lutte contre le sida, la tuberculose et le paludisme ont

été créées pour canaliser ces efforts internationaux. Il est alors devenu urgent de mesurer rigoureusement l'impact des interventions sur l'atteinte des objectifs et de disposer d'informations en vue de leur mise à l'échelle. L'intérêt porté à une approche de la santé mondiale fondée sur des preuves empiriques s'est ainsi accru, à l'instar de la révolution de la pratique clinique qui avait eu lieu quelques décennies plus tôt. Selon Cameron *et al.* (2016), 92,8 % de toutes les évaluations d'impact en matière de santé menées de 2000 à 2012, qui sont répertoriées dans le référentiel des évaluations d'impact (International Initiative for Impact Evaluation), étaient des RCT, la moyenne dans d'autres domaines du développement international étant de 66,4 % (Cameron *et al.*, 2016).

À la lumière de l'importance prise par les RCT dans le domaine de la santé mondiale, nous étudions ci-après les apports principaux des évaluations randomisées en la matière et mettons en évidence les limites qui pourraient être traitées de façon plus appropriée avec d'autres méthodes de recherche dans le cadre des politiques et des pratiques de santé mondiale.

Les apports en matière de politiques et de pratiques

Depuis l'an 2000, des centaines de RCT ont servi de base à l'élaboration de directives internationales dans des domaines prioritaires de la santé mondiale. Pour atteindre les OMD 4 et 5, les travaux d'évaluation en matière de santé maternelle et infantile se sont focalisés sur la réduction des retards de croissance des fœtus et des enfants, du dépérissement et des carences en micronutriments, qui comptent parmi les facteurs majeurs de mortalité infantile (Bhutta *et al.*, 2013). Les données issues de ces essais ont par exemple servi de base à la série d'actions recommandées par l'OMS (Organisation mondiale de la santé) et l'Unicef (United Nations International Children's Emergency Fund) pour la supplémentation en fer, en acide folique et en calcium des femmes enceintes, la promotion de l'allaitement maternel et l'apport de vitamine A et de zinc aux enfants (Bhutta *et al.*, 2013). Les RCT ont également fourni des données empiriques sur le développement des plateformes de prestation communautaire de soins visant à traiter les maladies infantiles courantes telles que la diarrhée, le paludisme et les infections respiratoires, et démontré que de telles approches peuvent réduire les taux de mortalité infantile (Hatt *et al.*, 2015 ; Whidden *et al.*, 2018). Bien que les OMD 4 et 5 n'aient pas été atteints, la mise à l'échelle des interventions évaluées par le biais des RCT au niveau des plateformes de soins primaires et de prestation communautaire a contribué à réduire le taux de mortalité des enfants de moins de cinq ans de 53 % (soit un passage de 90,6 à 42,5 décès pour 1 000 naissances vivantes) (You *et al.*, 2015), et le taux de mortalité maternelle de 43,9 % (de 385 à 216 décès pour 100 000 naissances vivantes) (Alkema *et al.*, 2016) entre 1990 et 2015.

La majorité des RCT liées à la santé, réalisées entre 2000 et 2013 dans les pays à revenus faibles et intermédiaires, consistaient essentiellement à évaluer des interventions biomédicales isolées pour la prévention, le diagnostic ou le traitement d'une maladie particulière (Cameron *et al.*, 2016 ; Kelaher *et al.*, 2016). Plus de 1 300 RCT ont ainsi été menées sur le VIH (virus d'immunodéficience humaine ou sida) (763 RCT), le paludisme (665 RCT) et la tuberculose (165 RCT) (Kelaher *et al.*, 2016). Elles ont fourni des preuves solides en faveur d'interventions directes de prévention du VIH, comme la prophylaxie antirétrovirale avant exposition, la circoncision médicale volontaire des hommes, ainsi que de nouveaux traitements efficaces (Krishnaratne *et al.*, 2016). En outre, les RCT relatives au paludisme ont permis de tester de nouveaux traitements pour les cas de paludisme simples, tels que les polythérapies à base d'artémisinine, ainsi que des approches préventives comme les moustiquaires imprégnées d'insecticide, le traitement préventif intermittent du paludisme pendant la grossesse ou la prophylaxie du paludisme chez les enfants (Bhutta *et al.*, 2013 ; Martinez-Alonso et Ramos, 2016). Les interventions communautaires contre la tuberculose associées à une thérapie sous observation directe se sont avérées très efficaces pour améliorer les taux d'observance et de réussite du traitement (Arshad *et al.*, 2014 ; South African Cochrane Centre, 2014), tandis que l'utilisation de nouveaux traitements préventifs s'est révélée efficace pour la prévention de la tuberculose chez les personnes infectées ou non par le VIH (South African Cochrane Centre, 2014). Les programmes actuels d'administration de masse de médicaments, qui constituent le pilier des stratégies de contrôle et d'élimination de nombreuses maladies tropicales négligées (par exemple les géohelminthiases transmises par le contact avec le sol, la filariose lymphatique ou la schistosomiase), ont été étendus à l'échelle internationale suite à des RCT (Kappagoda et Ioannidis, 2014). Parmi les exemples notables, on peut citer une RCT réalisée en 2004 par Kremer et Miguel, qui a montré les effets significatifs des médicaments antiparasitaires sur l'absentéisme et les performances scolaires au Kenya (Miguel et Kremer, 2004), et a conduit dans le monde entier à la mise en œuvre de programmes de déparasitage de portée nationale (Hatt *et al.*, 2014). Il ne s'agit là que de quelques exemples où les RCT ont permis à la communauté sanitaire mondiale de disposer d'un ensemble d'interventions pour réduire l'incidence des maladies. La mise à l'échelle de ces interventions (parmi de nombreux autres facteurs) a contribué à réduire la mortalité due au paludisme de 58 % et les nouvelles infections par le VIH de 40 % entre 1990 et 2015. On estime par ailleurs que 37 millions de décès dus à la tuberculose ont pu être évités au cours de la même période (United Nations, 2015).

Un certain nombre de RCT importantes ont également fourni des informations sur des politiques et réformes de santé publique plus larges avant leur mise en œuvre à l'échelle nationale ou internationale (Gertler, 2004). Un exemple classique est l'évaluation du programme mexicain Progresa, consistant en transferts monétaires pour permettre aux ménages pauvres de participer à une série d'activités liées à la santé comme les soins prénataux, la protection infantile, la vaccination, le suivi nutritionnel, ainsi que des programmes éducatifs de

promotion de la santé. Les essais, qui ont alloué ces aides de manière aléatoire à des groupes spécifiques, ont montré une réduction cohérente des taux de maladie dans les groupes ayant bénéficié de l'action par rapport à la population de contrôle (Gertler, 2004). Des programmes de transferts monétaires ont depuis lors été testés et mis en œuvre dans le monde entier (Hatt *et al.*, 2014). Les RCT ont également apporté des éclairages sur des réformes des soins de santé, comme les programmes de financement basé sur la performance (FBP). Une évaluation initialement randomisée de cette stratégie de FBP appliquée au Rwanda a démontré son impact, et fourni des données pour le déploiement national qui a été mené par la suite dans le cadre de la refonte du système de santé rwandais (Kruk *et al.*, 2016). Depuis, plus de 20 pays africains ont lancé ou commencé à intensifier des programmes FBP dans le domaine des soins de santé (Meessen *et al.*, 2011). Les résultats d'une série de RCT ont également nourri le débat sur l'introduction de copaiements d'un montant peu élevé pour des services préventifs et curatifs. Il est généralement admis que les copaiements sont essentiels pour promouvoir la durabilité, réduire les abus et garantir une utilisation judicieuse des produits et services (Bates *et al.*, 2012). Les RCT ne cessent toutefois de démontrer que le fait de facturer des frais peu élevés pour des produits préventifs tels que le savon, les moustiquaires, les vermifuges ou les agents de désinfection de l'eau réduit considérablement l'accès à ces produits pour ceux qui en ont le plus besoin, alors que cela ne génère que peu de recettes (Bates *et al.*, 2012). Bien que l'adoption généralisée de ces idées ait pris du retard, les gouvernements proposent aujourd'hui la plupart de ces produits à titre gratuit dans le cadre de leurs politiques nationales de santé.

Malgré le très grand nombre de preuves empiriques produites par les RCT dans des domaines clés de la santé mondiale, les approches sectorielles (qui présentent des avantages transversaux, mais sont intrinsèquement plus complexes) se prêtent moins aux RCT (Frieden, 2017 ; Deaton et Cartwright, 2018). Le fait de considérer la RCT comme un étalon-or universel en matière de méthodes d'évaluation d'impact et comme une condition préalable à l'élargissement des interventions peut avoir des conséquences inattendues sur les politiques de santé.

Des conséquences inattendues : un écart croissant en matière de preuves empiriques et de financement dans des secteurs clés de la santé

Jusqu'à l'établissement des OMD, il existait un équilibre relatif – et de nombreux débats – sur l'efficacité comparée des interventions sanitaires verticales et des interventions horizontales plus intégrées (visant des

« systèmes »). L'ère des OMD a rompu cet équilibre, la plupart des financements et des efforts étant concentrés sur un éventail de programmes verticaux. Du fait de l'urgence et de l'attention portée à quelques domaines prioritaires, les programmes verticaux ont été privilégiés, car ils étaient supposés permettre une plus grande spécialisation des services, accorder davantage de place aux maladies hautement prioritaires, accroître la responsabilisation, produire des résultats plus rapides et avoir de meilleures chances de réussite dans les États faibles (ATUN *et al.*, 2008). L'aide au développement en matière de santé a connu une croissance exponentielle au profit des programmes verticaux visant la santé des enfants (par exemple, la vaccination, la malnutrition), la santé maternelle, le VIH, le paludisme et la tuberculose, passant d'environ 3 à 4 milliards de dollars par an en 1990-2000 à plus de 24 milliards en 2016 (Institute for Health Metrics and Evaluation [IHME], 2016). Cette évolution a été soutenue en parallèle par l'augmentation des preuves empiriques issues des RCT pour la mise à l'échelle d'actions efficaces (voir section précédente).

Malgré leurs avantages, les programmes verticaux sont des approches descendantes pilotées de l'extérieur qui, en l'absence d'investissements parallèles dans des systèmes de santé plus solides, peuvent avoir des répercussions négatives comme la fragmentation des services, l'augmentation des obstacles à l'accès aux soins de santé pour les populations non ciblées et la réduction de l'efficacité et de la durabilité des systèmes de santé (ATUN *et al.*, 2008). Les interventions horizontales, comme le renforcement des systèmes de santé et les approches sectorielles, sont complexes par nature, nécessitent une adaptation spécifique au contexte et agissent à de multiples niveaux d'un système de santé (PLSEK et GREENHALGH, 2001 ; CAMPBELL *et al.*, 2007). Leur évaluation au moyen de RCT présente des difficultés importantes, nécessite des investissements conséquents, et s'avère bien souvent irréalisable (PLSEK et GREENHALGH, 2001 ; CAMPBELL *et al.*, 2007). Comme les actions de renforcement des systèmes de santé et les approches sectorielles souffrent à la fois d'un faible engagement politique et d'un manque de preuves d'efficacité fondées sur des RCT, le pourcentage de l'aide au développement pour la santé qui leur est alloué a diminué, passant d'environ 15 % en 1990 à moins de 10 % en 2016 (fig. 1) (Institute for Health Metrics and Evaluation [IHME], 2016). Comparées à la plupart des domaines d'intervention liés à la santé, les augmentations des financements au profit des actions de renforcement des systèmes de santé et des approches sectorielles n'ont pas suivi la tendance dominante. Les hausses annuelles moyennes des financements dans ce domaine sont passées de 11,4 % pour la période 1990-1999 à 7,1 % pour la période 2000-2009 (alors que tous les autres domaines enregistraient une hausse de ces taux de croissance). Avec une baisse en valeur absolue de 2,3 % subie entre 2010 et 2016, c'est l'un des seuls domaines d'action de la santé à avoir connu une telle évolution au cours de cette période (IHME, 2016).

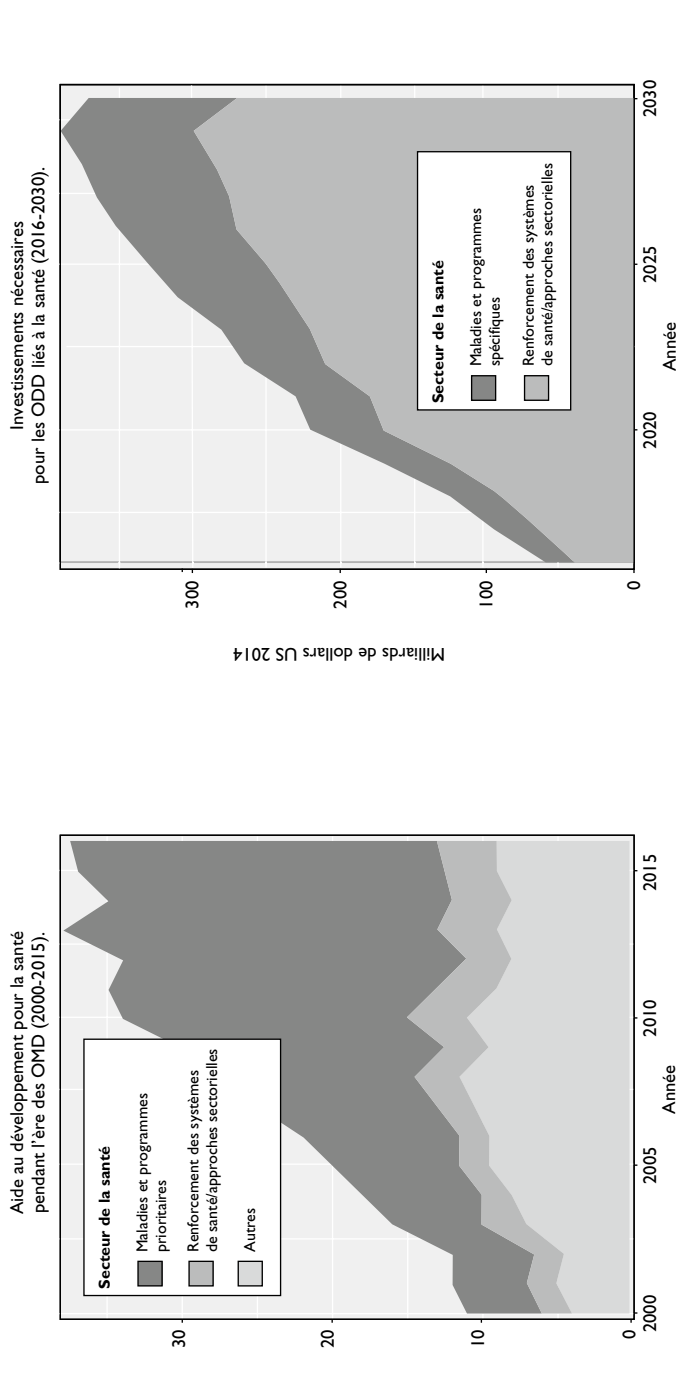


Figure 1

Évolution des investissements en matière de santé entre l'ère des Objectifs du millénaire pour le développement (OMD) (2000-2015) et l'ère des Objectifs de développement durable (ODD) (2016-2030).

Source : Andres Garchitorena, Megan Murray, Bethany Hedt-Gauthier, Paul Farmer et Matthew Bonds.

Note : la figure de gauche représente l'aide totale au développement allouée à la santé par an, adaptée des données de l'IHME (2016). La figure de droite, adaptée des données de Stemberg *et al.* (2017), représente la projection des investissements supplémentaires nécessaires dans 67 pays à revenus faibles et intermédiaires pour atteindre l'Objectif de développement durable (ODD) 3 lié à la santé. Elle montre qu'une augmentation considérable des financements pour le renforcement des systèmes de santé (RSS) et les approches sectorielles est nécessaire au cours des quinze prochaines années.

Le programme post-OMD des Nations unies, qui s'articule autour des Objectifs de développement durable (ODD), traduit une tentative pour réduire cet écart en se concentrant explicitement sur des approches sectorielles comme la couverture sanitaire universelle (CSU) et le renforcement des systèmes de santé. L'OMS a estimé que pour atteindre les ODD liés à la santé, près des trois quarts de l'ensemble des investissements supplémentaires requis pour les pays à revenus faibles et intermédiaires devraient être alloués aux mesures de renforcement des systèmes de santé et aux approches sectorielles au cours de la période 2015-2030, ce qui représente environ 300 milliards de dollars par an d'ici 2030 (Stenberg *et al.*, 2017) (fig. 1). Une évolution aussi radicale nécessite de repenser en profondeur les données nécessaires et les méthodes d'évaluation appropriées pour éclairer les décisions sur l'affectation et la mise en œuvre des financements. Les organisations internationales comme la Banque mondiale, l'OMS et l'United States Agency for International Development (USAID), entre autres, s'accordent de plus en plus à dire que la base de connaissances actuelle permettant d'orienter cette intégration horizontale est terriblement inadaptée, malgré ses avantages apparents (Atun *et al.*, 2008 ; Giedion *et al.*, 2013 ; Hatt *et al.*, 2015). Une revue des études systématiques sur le renforcement des systèmes de santé réalisée par l'USAID en 2015 a notamment conclu qu'« il est nécessaire de trouver des méthodes complémentaires pour estimer les effets des mesures de renforcement visant des systèmes adaptatifs complexes » (Hatt *et al.*, 2015).

Le manque de preuves empiriques dans des domaines clés de la santé tels que le renforcement des systèmes de santé et la couverture maladie universelle est symptomatique d'un problème plus large de la recherche dans le domaine de la santé : le décalage entre la portée des questions traitées par les RCT et le type de données nécessaires pour améliorer les résultats en matière de santé. On estime que 97 % des fonds de recherche sont consacrés au développement de nouvelles technologies de santé (essentiellement des produits pharmaceutiques), alors que 3 % seulement vont à la recherche sur la mise en œuvre (Kruk *et al.*, 2016). L'efficacité de la configuration des programmes et de l'optimisation de leur exécution est de ce fait rarement évaluée, ce qui entraîne des lacunes importantes pour la transposition des interventions en conditions réelles (Kruk *et al.*, 2016).

Défis et limites

L'une des questions les plus importantes en matière de santé mondiale est de savoir pourquoi des technologies connues – celles qui ont fait leurs preuves dans certains contextes – échouent systématiquement à toucher les personnes auxquelles elles sont destinées. La moitié de la population mondiale n'a pas accès à des services de santé essentiels (World Health Organization et World Bank, 2017). La majorité des décès d'enfants en Afrique subsaharienne est due à des maladies – diarrhée, paludisme, pneumonie – contre lesquelles les solutions

sont connues, peu coûteuses et efficaces. La thérapie de réhydratation orale, par exemple, permet d'éviter 90 % des décès d'enfants liés à la diarrhée dans le monde, mais seuls 4 enfants sur 10 qui en ont besoin en bénéficient (KRUK *et al.*, 2016). Dans la majorité des pays en développement, les ministères de la Santé ont établi des politiques nationales basées sur des normes internationales, mais, dans bien des cas, on ne sait pas quelle est la meilleure façon de les mettre en œuvre, même à petite échelle. Le défi est que même des technologies simples nécessitent des systèmes de prestation de soins complexes – personnels de santé formés, infrastructures, fournitures et médicaments – pour leur mise en œuvre sur le lieu de prise en charge. Les défaillances se manifestent à différents niveaux, depuis les agents de santé communautaires individuels aux établissements de soins de santé, en passant par les chaînes d'approvisionnement nationales, et se rétro-alimentent (BRUMMITT *et al.*, 2017). Ce phénomène a motivé le mouvement en faveur du « renforcement des systèmes de santé ». L'utilisation des RCT pour répondre à ces questions fondamentales comporte des difficultés intrinsèques. Pour un examen approfondi des limites méthodologiques des RCT, se reporter aux chap. 1 (Ravallion) et 2 (Pritchett) de ce volume, ou au débat suscité récemment par DEATON et CARTWRIGHT (2018). Nous présentons ci-dessous des sujets choisis relatifs à la pratique de la santé mondiale.

Le point fort d'une RCT bien menée est sa solide validité interne. Dans des conditions contrôlées, elle peut fournir une estimation non biaisée de l'effet de traitement moyen pour des interventions particulières, bien que les critiques affirment que cela n'est vrai que si de nombreuses hypothèses concernant la conception de l'essai sont confirmées (COOK, 2018). Même lorsque les résultats des RCT produisent une estimation non biaisée de l'effet pour la population spécifique étudiée, cela ne présage pas nécessairement l'effet de l'action sur d'autres populations en conditions réelles, ce qui signifie que l'on observe fréquemment des différences entre l'impact d'une action mise en œuvre dans des conditions contrôlées (« *efficacy* ») et son effet véritable dans le monde réel (« *effectiveness* ») (AHMED *et al.*, 2010 ; SHELTON, 2014). Contrairement aux études observationnelles, le processus de recrutement des participants à une RCT peut remplacer artificiellement les systèmes de prestation en place dans le monde réel afin de créer les conditions d'étude optimales, ce qui est particulièrement inutile car l'adhésion à un système de prestation performant constitue elle-même un problème central à résoudre. En outre, l'impact des interventions, tel que mesuré par les RCT réalisées dans des populations ou contextes différents, peut varier considérablement. Dans l'exemple du déparasitage cité plus haut, par exemple, les méta-analyses de plusieurs RCT ne permettent pas d'établir de manière concluante si les campagnes de masse ont un effet sur l'état nutritionnel, les résultats scolaires ou la survie des enfants, étant donné l'hétérogénéité des résultats des études et les différences dans les critères d'inclusion dans chaque méta-analyse (TAYLOR-ROBINSON *et al.*, 2015 ; CROKE *et al.*, 2016 ; VRIEZE, 2018).

Le coût élevé des RCT peut conduire les chercheurs à prendre en compte des périodes d'étude ou des tailles d'échantillons insuffisantes pour évaluer

correctement l'effet du traitement, ou à utiliser des indicateurs indirects qui ne correspondent pas au résultat évalué ou qui ne permettent pas de le déterminer (par exemple des indicateurs de processus, des signes limités ou précoces de maladie) (Frieden, 2017). À titre d'exemple, la suppression des paiements directs au point d'intervention par le biais des régimes d'assurance ou de dispenses de frais pour le bénéficiaire des services de soins constitue une stratégie clé pour accroître l'accès aux soins de santé, fournir une couverture financière contre des dépenses catastrophiques et améliorer finalement les résultats en matière de santé. Alors que la suppression des frais pour le bénéficiaire des soins peut impacter certains résultats sanitaires et économiques dans des populations de tous âges, une RCT réalisée au Ghana n'a mesuré que l'anémie entraînée par le paludisme chez les enfants de moins de cinq ans. L'essai a conclu que cette action n'avait pas d'effets mesurables sur la santé (Ansah *et al.*, 2009), mais l'évaluation a été effectuée six mois seulement après le début de l'intervention, et l'étude était sous-dimensionnée compte tenu de la faible prévalence de l'anémie observée (Ridde et Haddad, 2009). Les problèmes de conception tels que ceux illustrés dans cet exemple, bien qu'ils ne soient pas l'apanage des seules RCT, sont relativement courants. Une revue des RCT indexées dans PubMed publiées en 2001 et 2006 a révélé que de nombreuses évaluations étaient menées avec des tailles d'échantillons insuffisantes pour détecter même des effets de traitement importants (Hopewell *et al.*, 2010). Des conceptions d'études inadéquates, entre autres facteurs, contribuent à un gaspillage important des investissements dans la recherche biomédicale, estimé à environ 85 % des ressources investies (200 milliards de dollars en 2010) (Macleod *et al.*, 2014).

Certains chercheurs privilégient les RCT parce que la randomisation permet d'équilibrer les facteurs connus et inconnus qui influencent le résultat évalué, ce qui simplifie l'inférence statistique dans l'évaluation avec une connaissance minimale des mécanismes qui sous-tendent les effets observés. Ce modèle peut toutefois avoir un effet pervers car il réduit le flux d'informations entre les chercheurs et le contexte dans lequel ils réalisent les essais, facteur particulièrement important lorsque ce contexte implique des populations vulnérables agissant dans des conditions complexes. On peut également aboutir à des conceptions d'études non éthiques (Deaton et Cartwright, 2018). Les essais expérimentaux éthiques exigent par exemple que l'intervention soit réalisée selon le principe d'équipoise – c'est-à-dire avec une incertitude quant à ses bénéfices –, mais de nombreuses RCT sont effectuées dans le but de confirmer les résultats d'études observationnelles, en évitant d'offrir les avantages du programme aux sujets présents dans le groupe de contrôle (Farmer *et al.*, 2013) (pour plus de détails sur l'équipoise et ses implications pour les RCT, voir Abramowicz et Szafarz, chap. 10, ce volume). Quelle que soit la conception de l'étude, la création d'un ensemble de preuves empiriques dans le domaine des prestations de services en santé mondiale requiert l'existence de larges boucles d'information entretenues sur le long terme, dans lesquelles les acteurs locaux, les praticiens et les exécutants participent activement à la hiérarchisation des problématiques de recherche, ainsi qu'à l'interprétation et à la diffusion des résultats, générant

des possibilités de formation et de recherche encadrées qui apportent des éclairages sur les services fournis (Farmer *et al.*, 2013). Dans la section suivante, nous illustrons la façon dont des cadres d'évaluation complémentaires peuvent contribuer à la constitution de cet ensemble de preuves.

Au-delà des RCT pour les ODD : des cadres d'évaluation observationnelle pour le renforcement des systèmes de santé et la CSU

La recherche sur la mise en œuvre a pour but d'établir si, comment, quand et pourquoi une intervention fonctionne, et de proposer d'autres hypothèses ultérieures (Bhattacharyya *et al.*, 2009 ; Kruk *et al.*, 2016). Elle fait appel à divers modèles d'études, allant des méthodes observationnelles et expérimentales quantitatives à la recherche qualitative (Kruk *et al.*, 2016). L'utilisation des résultats d'études observationnelles, qui viennent parfois contredire leurs homologues expérimentales, fait l'objet de nombreux débats (Ioannidis *et al.*, 2001 ; Prasad *et al.*, 2013 ; Hemkens *et al.*, 2016 ; Jones et Steel, 2018). De multiples études comparant les essais randomisés et non randomisés montrent cependant que les études observationnelles de grande qualité (comme les études prospectives contrôlées) peuvent produire des résultats comparables à ceux des RCT (Ioannidis *et al.*, 2001 ; Jones et Steel, 2018). Le principal défi des études observationnelles est le risque accru de choisir des groupes de comparaison inappropriés avec des facteurs non mesurés qui peuvent fausser les résultats. Une évaluation initiale du projet « Villages du millénaire » a notamment été critiquée pour avoir choisi rétroactivement un contrôle biaisé qui avait orienté favorablement les résultats de l'étude (Mitchell *et al.*, 2018). Dans ce cas, les systèmes de données et d'évaluation étaient initialement insuffisants en l'absence de groupes de contrôle *a priori*, ce qui a compromis la capacité à tirer des conclusions définitives sur l'efficacité de l'intervention. Les analyses rétrospectives de suivi se sont toutefois avérées plus solides (Mitchell *et al.*, 2018). On note également que, dans certains cas, les méthodes observationnelles ont été préférées aux résultats des RCT pour la prise de décision, malgré leur divergence, parce qu'elles permettent des périodes de suivi plus longues, des tailles d'échantillon plus importantes et une probabilité plus forte de détecter des effets négatifs (Frieden, 2017). Les recommandations concernant la vaccination antigrippale par pulvérisation nasale avec des vaccins vivants atténués, qui avaient initialement démontré une bonne protection dans des RCT, ont par exemple évolué dans le temps après que des études observationnelles ultérieures ont suggéré que la validité externe des résultats des RCT était limitée (Frieden, 2017).

Concernant les approches sectorielles, le Medical Research Council britannique reconnaît que la conception, la description et la mise en œuvre d'une intervention complexe sont les points faibles les plus fréquents des RCT (CAMPBELL *et al.*, 2000). Il ne propose pas de méthodologies alternatives, mais fournit des lignes directrices explicites permettant la réalisation de RCT bien conçues pour des interventions complexes (CAMPBELL *et al.*, 2000). Il existe des exemples convaincants d'interventions complexes qui ont été évaluées par le biais de RCT (BANERJEE, *et al.*, 2015a), comme celles décrites ci-dessus pour le Mexique et le Rwanda. Les ressources nécessaires pour mener de telles évaluations limitent toutefois leur capacité à être utilisées à grande échelle dans les pays à revenus faibles et intermédiaires. Parallèlement, les chercheurs tentent de tirer des enseignements issus du domaine des « sciences de la complexité » pour mieux comprendre les systèmes de soins de santé, qui sont à la fois complexes et adaptatifs (PLSEK et GREENHALGH, 2001). La théorie de la complexité suggère qu'au lieu de décomposer le système en éléments simples (comme avec les RCT portant sur de multiples interventions verticales), il peut être préférable de mettre en œuvre simultanément diverses approches, et de s'orienter progressivement vers celle(s) qui fonctionne(nt) (équivalent à une mise en œuvre adaptative avec des études observationnelles quasi expérimentales) (PLSEK et GREENHALGH, 2001). Pour atteindre les objectifs de développement liés à la santé, comme la santé maternelle et infantile, le facteur le plus important peut être l'effet collectif d'une série optimale d'actions destinées à des populations et à des contextes particuliers (SHELTON, 2014). Dans cette optique, cumuler, en parallèle à la myriade d'interventions de prestation de soins de santé qui ont lieu dans les pays en développement, des méthodes solides de collecte de données et d'évaluation (notamment des méthodes observationnelles quasi expérimentales), représente l'une des opportunités majeures en matière de santé mondiale. Cela peut permettre de mener des recherches rigoureuses à moindre coût et sans devoir contrôler le processus de mise en œuvre ou la population bénéficiaire.

Les limites des RCT pour l'évaluation de programmes de santé mondiale complexes déployés à grande échelle ont conduit au développement de cadres qui prennent en compte, parallèlement à la mise en œuvre des programmes, un ensemble de méthodes observationnelles, telle que celles proposés par le Partenariat international pour la santé (IHP+), l'Initiative pour la santé en Afrique et l'Initiative catalytique pour sauver un million de vies (World Health Organization, 2010 ; VICTORA *et al.*, 2011 ; BRYCE *et al.*, 2013), auxquels participent des acteurs majeurs comme l'OMS, la Banque mondiale et la fondation Bill et Melinda Gates. Ces cadres d'évaluation déterminent la réussite des programmes en termes de gains de couverture des interventions et d'effets sur la santé. Les études sont réalisées en conditions réelles avec une mise en œuvre plus variable que dans les essais contrôlés. Les interventions se déroulent rarement de manière isolée, étant donné que de nombreuses organisations mettent en œuvre des programmes pratiquement partout dans les pays en développement, et que des changements s'opèrent au niveau des situations sanitaires et

socio-économiques indépendamment des programmes existants (World Health Organization, 2010 ; Victora *et al.*, 2011 ; Bryce *et al.*, 2013 ; El-Sadr *et al.*, 2014 ; Reidy *et al.*, 2018).

En se servant du district de santé comme unité d'étude, les chercheurs évaluent des indicateurs clés au niveau des intrants, des processus et des extrants du système de santé (par exemple, le personnel de santé, les services disponibles), conjointement à des indicateurs de résultat et d'impact (par exemple, la couverture des services et les taux de mortalité ; tabl. 1). Outre le suivi continu de la mise en œuvre du programme, la collecte de données complémentaires permet aux chercheurs de combler les lacunes d'information avant, pendant et après la période d'évaluation, en utilisant des évaluations des établissements de santé, des enquêtes auprès des ménages, des recherches longitudinales et des études qualitatives. Les analyses quantitatives sont complétées par des descriptions qualitatives de la mise en œuvre des programmes (c'est-à-dire ce qui est mis en œuvre et de quelle manière) et des facteurs contextuels qui peuvent avoir eu une incidence sur la mise en œuvre et l'impact, de façon à ce que les résultats puissent être interprétés de manière appropriée et des enseignements tirés (Victora *et al.*, 2011 ; Requejo *et al.*, 2015 ; Reidy *et al.*, 2018).

Au niveau national, ces évaluations à grande échelle de l'efficacité des programmes ont conduit à l'initiative « *Countdown* », qui assure le suivi d'une liste complète des indicateurs mentionnés ci-dessus pour chaque pays à revenus faibles ou intermédiaires, et permet ainsi des comparaisons objectives et solides des progrès réalisés par chaque pays (Requejo *et al.*, 2015). Au niveau infranational, ce cadre sert à évaluer l'impact des actions complexes de renforcement des systèmes de santé en contribuant à combler le manque important de preuves empiriques en la matière. Prenons à titre d'illustration deux expériences récentes menées au Rwanda (Thomson *et al.*, 2018) et à Madagascar (Garchitorena *et al.*, 2018), qui mettent en œuvre une série similaire d'actions de renforcement du système de santé intégrées à divers niveaux de prestations de soins (santé communautaire, centres de soins de santé primaires, hôpital de référence). Les deux interventions visent à améliorer la capacité du système de santé par le biais de programmes horizontaux tout en intégrant verticalement des programmes cliniques prioritaires. Des évaluations sont effectuées à l'aide des données recueillies au niveau des ménages dans le cadre d'enquêtes démographiques et de santé transversales, et émanant d'échantillons représentatifs des populations respectives (pour Madagascar, la démarche incluait une enquête de base représentative à la fois de la zone de couverture de l'intervention et de la zone de comparaison), et d'enquêtes répétées à intervalles rapprochés pendant toute la durée des interventions (tous les cinq ans pour le Rwanda et tous les deux ans pour Madagascar). L'impact des interventions est évalué au moyen d'analyses statistiques similaires à la méthode des différences de différences pour un large éventail d'indicateurs de résultat du type de ceux présentés dans le tabl. 1, en contrôlant par des facteurs confondants pertinents (par exemple, la richesse des ménages).

Tableau 1
Indicateurs de couverture sanitaire et de mortalité dans le continuum de soins pour la santé maternelle et infantile.

Indicateurs de couverture sanitaire (%)	**Indicateurs de mortalité**
Prégrossesse Demande de planification familiale satisfaite	
Grossesse	
Suivi prénatal (⩾ 1 visite) Suivi prénatal (⩾ 4 visites) Traitements préventifs intermittents du paludisme durant la grossesse Protection contre le tétanos néonatal	
Naissance Accouchement assisté par du personnel soignant qualifié	Mortalité maternelle (nombre de décès pour 100 000 femmes)
Période postnatale Visite postnatale pour les mères Visites postnatales pour les nourrissons Initiation précoce à l'allaitement maternel	Mortalité néonatale (nombre de décès pour 1 000 naissances vivantes)
Petite enfance Allaitement maternel exclusif (< 6 mois) Introduction des aliments (6-8 mois) Vaccination DTCoq (diphtérie-tétanos-coqueluche), 3 doses Première injection du vaccin contre la rougeole Vaccination Hib3 (hæmophilus influenzæ de type b) Supplémentation en vitamine A (2 doses)	Mortalité infantile (nombre de décès pour 1 000 naissances vivantes)
Enfance Enfants dormant sous des moustiquaires imprégnées d'insecticide Recours aux soins pour les symptômes de la pneumonie Traitement antipaludéen de première intention Traitement aux sels de réhydratation orale Sources d'eau potable améliorées Installations sanitaires améliorées	Mortalité des enfants de moins de cinq ans (nombre de décès pour 1 000 naissances vivantes)
Indicateurs composites de santé materno-infantile Indice de couverture composite (ICC)	

Source : Andres Garchitorena, Megan Murray, Bethany Hedt-Gauthier, Paul Farmer et Matthew Bonds, adaptation à partir des données de Requejo *et al.* (2015).

Ce modèle d'étude quasi expérimental permet aux responsables de programmes de disposer de l'autorité nécessaire sur la mise en œuvre des programmes (quand, où et comment les actions sont déployées) qui n'est pas prescrite par un protocole de recherche. Des systèmes de données sont construits autour de l'intervention de façon à ce que les chercheurs puissent évaluer les activités en cours et fournir des informations susceptibles d'aider les responsables à adapter les programmes sans perturber la mise en œuvre. Une analyse réalisée sur la période 2014-2016

à Madagascar a par exemple montré que, malgré des améliorations globales de la plupart des indicateurs de couverture, l'accès aux soins de santé restait très faible pour les populations éloignées des établissements de santé (Garchitorena *et al.*, 2018). Ce constat a entraîné une expansion du soutien aux programmes de santé communautaire, tant géographiquement que dans la portée des services fournis, sans que la conception de l'étude ne s'y oppose. En outre, des études complémentaires de recherche sur la mise en œuvre permettent d'évaluer des composants spécifiques au sein de l'intervention globale, comme un programme de mentorat et de supervision accrue, mené dans le cadre de l'intervention de renforcement du système de santé au Rwanda (Manzi, *et al.*, 2018a ; 2018b), tout en contribuant à améliorer les capacités de recherche au niveau des praticiens locaux (Hedt-Gauthier *et al.*, 2017 ; Odhiambo *et al.*, 2017).

Suite à l'intervention de renforcement du système de santé dans un district et demi du Rwanda rural, la mortalité des enfants de moins de cinq ans a baissé de plus de 60 % entre 2005 et 2010 (Thomson *et al.*, 2018). Cette réduction était beaucoup plus importante que celle atteinte dans le reste du pays, et trois fois supérieure au taux requis pour satisfaire les OMD. De même, la mortalité des enfants de moins de cinq ans et la mortalité néonatale ont baissé respectivement de près de 20 % et 35 % au cours des deux premières années de l'intervention de renforcement du système de santé dans un district de Madagascar (2014-2016), à un rythme significativement plus rapide que les taux nationaux moyens observés pour n'importe quel pays pendant la période des OMD. Même si les caractéristiques de référence étaient similaires en termes de revenu par habitant et de taux de mortalité des enfants de moins de cinq ans, chaque intervention s'est déroulée dans des contextes politiques et économiques très différents (Bonds et Rich, 2018). Au cours de la période 2005-2010, le Rwanda a connu un cycle vertueux de stabilité politique, d'investissements internationaux et d'aide étrangère. À Madagascar, en revanche, la situation politique a été instable pendant la majeure partie des 50 dernières années, avec une économie en déclin constant et les investissements dans le système de santé les plus faibles à l'échelle mondiale en 2014. À elles deux, ces expériences constituent un test naturel de la portée des impacts que les interventions intégrées de renforcement des systèmes de santé peuvent avoir au niveau de la population, et qui peuvent être reproduits dans des contextes différents (Bonds et Rich, 2018).

Conclusion

Les deux dernières décennies ont vu une amélioration sans précédent des indicateurs de santé dans les pays à revenus faibles et intermédiaires, largement liée aux OMD. Les RCT ont joué un rôle essentiel au cours de cette période, en facilitant l'adoption de technologies et de services médicaux efficaces pour réduire l'incidence des maladies. La plupart de ces services ont fait l'objet de

tentatives de mise à l'échelle verticale, qui peut elle-même se prêter à des RCT. Pourtant, d'importantes régions du monde continuent à souffrir d'un manque d'accès aux soins de santé primaires. Les échecs rencontrés dans le processus de mise à l'échelle sont souvent dus à des défaillances dans la mise en œuvre de base et à la faiblesse des systèmes de santé. Un consensus se fait ainsi autour de l'importance centrale des approches sectorielles comme le renforcement des systèmes de santé intégrés, les soins primaires intégrés et la couverture sanitaire universelle. Pour atteindre les objectifs de développement durable liés à la santé (ODD 3), les investissements dans ces secteurs devront être multipliés par plus de cinq au cours des 15 prochaines années. Quel type de preuves empiriques devrait guider ces investissements et éclairer la mise en œuvre ?

Les approches sectorielles se prêtent moins bien aux RCT, et les modèles d'études devraient être guidés par les priorités programmatiques. Si les RCT sont toujours fondamentales pour déterminer des solutions isolées et pertinentes dans un large éventail de contextes, la recherche sur la mise en œuvre, qui fait généralement appel à des méthodes à la fois qualitatives et quantitatives sans nécessairement randomiser la mise en œuvre, peut aider les responsables des programmes à comprendre comment des actions dont les effets ont été démontrés peuvent être intégrées efficacement dans les systèmes de prestations de soins de santé. Les méthodes observationnelles et quasi expérimentales, en particulier, sont les plus appropriées lorsque l'échelle de l'intervention rend la randomisation impossible ou difficilement réalisable. Le fait d'adjoindre à la myriade d'interventions de prestations de soins de santé organisées dans les pays en développement une collecte de données solide pourrait aider au progrès d'une recherche rigoureuse à faible coût sans nécessité de contrôler la mise en œuvre ou la population bénéficiaire. Un examen complet des différents types de preuves disponibles pourrait contribuer à orienter les efforts déployés en matière de santé mondiale au cours des prochaines décennies.

Chapitre 6

Essais et tribulations

L'essor et le déclin des expérimentations aléatoires dans le secteur de l'eau, de l'assainissement et de l'hygiène

Dean Spears, Radu Ban et Oliver Cumming

Introduction : le besoin de réflexion

Que gagne-t-on, ou pas, à se concentrer sur les résultats des études randomisées ? Perd-on quelque chose ? Ces derniers temps, au sein de la communauté internationale du développement, ceux qui s'intéressent au secteur de l'eau, de l'assainissement et de l'hygiène (*Water, Sanitation and Hygiene* – WASH) se sont autant passionnés pour ces questions que ceux qui interrogent « l'efficacité » du développement dans son ensemble. Dans ce chapitre, nous présentons nos propres « observations participantes » sur ce débat en cours. Notre conclusion est que les évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) peuvent constituer une forme importante de preuves empiriques, au même titre que les études observationnelles. Mais si les RCT proposent effectivement des calculs clairs et simples, les essais randomisés WASH n'ont pas permis d'arriver à des conclusions plus précises que les autres méthodes. Notre expérience concorde avec les enseignements tirés par Deaton et Cartwright (2018) du recours aux RCT dans le domaine du développement : la randomisation « ne nous dispense pas de la nécessité de réfléchir ». Comme c'est aussi le cas, bien évidemment, pour tout autre type de preuve empirique.

La raison principale pour laquelle cette réflexion est requise est que personne ne peut raisonnablement s'attendre à ce que des RCT différentes menées dans le secteur WASH donnent les mêmes résultats dans des contextes différents et avec des caractéristiques distinctes en termes de maladies, de démographie, de culture et d'environnement, ou même en testant des types d'intervention différents.

L'hétérogénéité des résultats n'a rien de surprenant au regard de celle des données d'entrée. Il en résulte que l'interprétation des RCT, même si elles sont de grande qualité, nécessitera toujours une réflexion tout aussi qualitative que ce qu'exigent les méthodes observationnelles sans randomisation, ainsi que des connaissances théoriques. En outre, certains sujets indéniablement pertinents sur le plan des politiques ne se prêteront jamais à une expérimentation randomisée. Ainsi, ni les conduites d'égouts, ni la religion, ni la densité de population ne feront l'objet d'une randomisation (même si cela peut être le cas de traitements qui interagissent avec ces éléments). Et même certaines questions qui seraient en principe susceptibles d'être soumises à des traitements randomisés ne le seront probablement pas dans la décennie qui reste à courir avant l'échéance de l'Objectif de développement durable lié à l'assainissement.

Tout au long de cette partie, nous invitons le lecteur à comparer et opposer une RCT appliquée à une intervention WASH dans une zone rurale d'un pays en développement au cas paradigmatique en matière de RCT : un essai clinique de médicament dans une recherche médicale. Lors d'essais répétés sur un même médicament, les chercheurs testent en effet sur le plan chimique *le même médicament*. Et en fin de compte, si des séries d'essais de médicaments sont mises en œuvre de manière comparable, elles fournissent des estimations multiples du même paramètre causal, qui se prête à une méta-analyse. Une même molécule chimique n'a pas exactement le même effet dans chaque organisme, mais les résultats entre différentes études proviendront probablement de la même distribution. Même si les populations soumises aux essais cliniques diffèrent, les théories et les preuves issues de la littérature médicale seront certainement assez riches pour suggérer des explications. Les essais WASH, en revanche, sont inévitablement peu nombreux et différents.

Dans ce chapitre, faute de place nous ne pouvons couvrir tous les aspects importants des problématiques WASH dans les pays en développement. Les infrastructures urbaines, comme les systèmes d'égouts ou de traitement des matières fécales, seraient quasiment impossibles à évaluer dans le cadre d'une étude randomisée, mais elles sont toujours considérées comme un facteur important en matière de santé publique, si l'on s'appuie notamment sur la littérature relative à la démographie (Cutler et Miller, 2005[1]). Une vaste littérature dans le secteur WASH, que nous n'aborderons pas, questionne le rôle de la « mise en aveugle » dans les RCT menées à l'échelle des foyers. Au niveau de l'hygiène des mains, par exemple, on peut dissimuler si un savon est antibactérien ou non, mais pas si des membres du foyer sont encouragés à se laver les mains. Une étude influente sur le traitement de l'eau dans les foyers a permis de noter le rôle joué par l'intégration de la mise en aveugle dans la conception physique des filtres à eau. Dans le reste de ce chapitre, nous avons mis ces questions de côté pour nous concentrer sur l'*assainissement en milieu rural* dans les pays

1. Bien que cela ne change pas substantiellement notre argumentation, il convient de noter qu'Anderson *et al.* (2018) réévaluent les résultats de Cutler et Miller et constatent que les effets de la filtration de l'eau sont nettement plus faibles que dans la publication originale.

en développement, non pas parce que c'est la seule dimension importante de la problématique WASH, mais parce que c'est un cas utile pour comprendre l'apport potentiel des RCT.

Contexte : les preuves empiriques récentes en matière d'assainissement et de santé infantile

Le secteur de l'assainissement en milieu rural fait l'objet d'une certaine confusion et de nombreux débats. L'ambiguïté et la contestation ne sont en elles-mêmes pas surprenantes. Ce qui l'est davantage dans ce cas, c'est que les désaccords et la perplexité soient en grande partie *la conséquence de RCT récentes*, alors que celles-ci sont censées apporter une certaine clarté. Trois RCT sur l'assainissement réalisées il y a peu de temps (Luby *et al.*, 2018 ; Null *et al.*, 2018 ; Humphrey *et al.*, 2019) ont en particulier suscité un vif débat, tant dans les cercles scientifiques que dans les organisations du secteur de l'assainissement qui mettent en œuvre des programmes. Une publication de consensus (Cumming *et al.*, 2019) présente une interprétation commune des résultats émanant d'un groupe multidisciplinaire de chercheurs. Comme nous l'expliquons plus en détail ci-après, ces trois essais partagent le fait qu'ils aboutissent tous à des résultats nuls sur l'impact que peuvent avoir des systèmes sanitaires améliorés sur le rapport taille-âge chez les enfants. Ce qui rend ce résultat encore plus paradoxal en apparence, c'est que de nombreux observateurs jugent que les trois RCT considérées ont donné globalement la même réponse (le débat porte sur la juste conclusion à en tirer). En outre, personne ne prétend que les essais n'ont pas été menés avec soin, ou qu'ils n'ont pas atteint leur objectif.

Qu'est-ce qui a mal tourné ? Bien que la situation fâcheuse dans laquelle se trouve actuellement le secteur WASH ne soit pas celle imaginée par les partisans des RCT, elle concorde avec certaines des mises en garde de Cartwright et Deaton. En particulier, même s'il y a un large consensus sur *la nature* de l'effet du traitement qui a été évalué par les RCT, il y a un désaccord général sur le « *pourquoi* ». Il y a davantage d'explications potentielles qu'il n'y a d'éléments d'information fournis par les RCT. Cette question – le « pourquoi » – a besoin d'être posée pour élaborer des politiques : quels sont les facteurs importants à prendre en compte la prochaine fois. Mais, comme le soulignent Cartwright et Deaton, ce « pourquoi ? » est exactement la question à laquelle les RCT peuvent avoir du mal à répondre (Deaton et Cartwright, 2018).

Dans ce chapitre, nous examinons les raisons pour lesquelles ce résultat apparemment paradoxal ne doit pas surprendre : les problématiques WASH dans les pays en développement comportent en fait de nombreuses caractéristiques qui rendent les effets hétérogènes et dépendants du contexte. Par conséquent, une petite série d'études ne peut permettre de répondre à ce défi. Et c'est là un problème majeur quand on sait que de telles études coûtent des dizaines de millions de dollars et peuvent prendre des années avant d'être achevées. Mais avant de pouvoir tirer des leçons aussi générales, nous présentons dans le paragraphe qui suit les preuves empiriques issues de ces études récentes, ainsi que d'autres recherches.

Les enseignements tirés d'une série d'études

L'exposition à de mauvaises conditions sanitaires est largement soupçonnée d'être néfaste pour la santé des enfants. Il semble intuitivement plausible à la plupart des individus qu'un enfant qui est régulièrement exposé à des matières fécales puisse être, en moyenne, dans une situation pire qu'un enfant qui ne l'est pas. Mais à quel point sa situation est-elle pire, et quels sont les remèdes qui fonctionnent suffisamment bien ? Pour mesurer à quel point la situation est pire, il faut notamment choisir un indicateur de résultat. L'indicateur le plus important est sans doute la mortalité précoce, mais les résultats binaires ont une faible puissance statistique et, heureusement, même dans les populations les plus défavorisées, la plupart des enfants ne meurent pas. Il en résulte que l'étude de la mortalité nécessite des échantillons très vastes, trop vastes pour pouvoir mener une RCT crédible. Un autre indicateur est la diarrhée : les chercheurs ont interrogé des mères sur la mollesse ou la dureté des selles de leurs enfants dans des centaines d'enquêtes. Le défi en la matière est que dire d'une selle qu'elle est molle est un jugement subjectif : en Inde, des mères plus instruites signalent davantage de diarrhées, probablement parce qu'elles sont plus à même de considérer une situation donnée comme un problème.

Une variable de résultat statistiquement idéale a émergé au cours de la dernière décennie. Il s'agit d'une variable continue, mesurée objectivement et liée à une caractéristique que tous les individus possèdent : l'anthropométrie, et en particulier la taille des enfants. L'hypothèse que la taille peut être influencée par l'exposition à de mauvaises conditions sanitaires a gagné en crédibilité auprès des chercheurs du secteur WASH (Humphrey, 2009) en même temps qu'une découverte des économistes sur la taille des individus : son importance comme critère du capital économique humain pour ce qu'elle révèle sur la santé, la maladie et la nutrition dans l'enfance (Case et Paxson, 2008). Dans certains pays, comme dans le cas du mouvement politique indien en faveur du « droit à l'alimentation », les statistiques sur la taille des enfants ont même fait l'objet d'une grande attention dans les discussions sur les politiques publiques, en tant que mesure des résultats et des privations nutritionnels. Les acteurs politiques associés aux partis de gauche et de droite en Inde ont débattu de la question de savoir si la petite taille moyenne des enfants indiens relevait de facteurs génétiques non problématiques, de déficits alimentaires prioritaires, du statut des femmes, ou d'une exposition généralisée à la défécation en plein air (Coffey *et al.*, 2013).

Ceci a donné lieu à la publication d'une série d'articles étudiant l'effet des différentes dimensions de l'assainissement, notamment sur la taille. Les notions exactes étudiées en matière d'« assainissement » sont très variables : certains considèrent la défécation en plein air sans utilisation de toilettes ou de latrines ; d'autres étudient l'utilisation de latrines améliorées, plutôt que de latrines plus simples ; d'autres encore pensent qu'il est important de prendre en compte les moyennes relatives aux conditions sanitaires à l'échelle de la communauté, plutôt que le comportement des foyers isolés. Ces études peuvent globalement

être classées en quatre groupes, selon la méthodologie utilisée (RCT ou étude observationnelle) et la qualité (impact plus élevé ou plus faible), conduisant ainsi à des RCT et études observationnelles de qualité supérieure et inférieure.

Les RCT à faible impact

Un premier groupe de RCT n'a pas eu autant d'impact sur la réflexion dans ce secteur que les expérimentations plus récentes. Leur mise en œuvre a parfois pâti d'irrégularités ou d'un manque d'exhaustivité qui ont limité les leçons susceptibles d'en être tirées. Certaines ont été organisées dans plusieurs pays par l'intermédiaire du programme pour l'eau et l'assainissement de la Banque mondiale. L'une de ces études, dans le Maharashtra, n'a finalement pas été réalisée par le gouvernement de l'État dans tous les districts prévus par la Banque mondiale (Hammer et Spears, 2016). Dans une autre, menée dans l'État de Madhya Pradesh, le traitement relatif à l'assainissement a été mis en œuvre tellement tard qu'il ne restait plus que quelques semaines avant l'évaluation finale (Patil *et al.*, 2014). Une autre RCT a été effectuée dans l'Odisha (anciennement Orissa) en respectant les niveaux d'attention et de rigueur les plus élevés qui soient, mais n'a pas été en mesure de fournir beaucoup d'informations sur l'effet de la défécation en plein air sur les résultats infantiles, car l'intervention n'a en fin de compte pas donné lieu à une différence notable dans la défécation en plein air entre le groupe de traitement et le groupe de contrôle (Clasen *et al.*, 2014). Comme nous le verrons, compte tenu de normes sociales spécifiques, la promotion des latrines en Inde rurale, en substitution à la défécation en plein air, répond à des enjeux également spécifiques. Dans d'autres pays en développement, la défécation en plein air tend à devenir de plus en plus rare.

Les études observationnelles à faible impact

Les publications les plus nombreuses sont peut-être les multiples études observationnelles qui donnent peu d'indications crédibles, en particulier sur l'effet causal que pourrait avoir l'assainissement. Dans de nombreux cas, bien sûr, l'objectif de l'étude consistait simplement à décrire une situation ou une tendance émergeant des données disponibles. On ne peut pas reprocher à ces études de ne pas donner ce qu'une RCT est censée donner. Un grand nombre d'entre elles ont toutefois trop tendance à tirer des conclusions causales à partir de comparaisons (peut-être avec des contrôles de régression), alors que leurs caractéristiques ne permettent en fait pas d'assimiler corrélation et causalité. Certaines de ces études reviennent simplement à comparer la taille d'enfants vivant dans des foyers qui disposent de toilettes ou de latrines à celle d'enfants vivant dans un foyer qui n'en dispose pas. Nous ne connaissons aucun cas dans lequel une telle comparaison des conditions sanitaires au niveau des foyers, basée sur l'observation, offrirait une évaluation crédible d'un effet de causalité, ne serait-ce que parce que l'on ne prend pas en compte les effets d'entraînement de certains foyers d'une localité sur leur voisinage.

Beaucoup de chercheurs menant des études observationnelles sont donc suffisamment prudents pour ne pas établir d'inférences causales à partir de comparaisons

de ce type, mais ce n'est pas le cas de tous. Ce qui importe ici, c'est de savoir de quel côté les auteurs se situent, c'est-à-dire s'ils sont clairs et prudents au sujet des questions posées, ou s'ils exagèrent le résultat des données qu'ils ont obtenues (que ce soit dans l'étude, ou dans le cadre de discussions plus larges sur les politiques).

Les études observationnelles à fort impact

Un troisième groupe d'études est également de nature observationnelle, mais prend en compte avec plus d'attention les preuves empiriques susceptibles de fournir des conclusions véritablement instructives. Nombre d'entre elles s'inspirent de ce que certains chercheurs empiriques en économétrie ont appelé la « révolution de la crédibilité » des études d'identification causale. Les méthodes minutieuses d'investigation, de mise en doute et de double vérification des preuves d'impact issues de données d'observation sont néanmoins bien antérieures aux dernières décennies de la recherche économétrique. En effet, certains experts, comme le statisticien social FREEDMAN (1991), font remonter leur origine à l'étude de John Snow sur le secteur WASH et le choléra à Londres. Des démographes, des épidémiologistes, au même titre que des économétriciens, ont tous participé aux meilleures publications dans ce domaine.

Ce qui distingue certaines de ces études, c'est qu'elles recherchent des cas particuliers dont on puisse tirer des leçons. Ceci implique généralement de comprendre en profondeur et de manière spécifique au contexte les raisons pour lesquelles des facteurs susceptibles d'entraîner une confusion dans d'autres cas ne poseront probablement aucun problème dans le cas considéré. Dans l'étude originale de SNOW (1855) par exemple, celui-ci affirme que le choléra est transmis par la contamination du système de distribution de l'eau par des matières fécales. Son argument est basé sur les taux de mortalité liée au choléra beaucoup plus élevés parmi les foyers approvisionnés par la compagnie des eaux Southwark et Vauxhall, dont la prise d'eau de la Tamise était située en aval du point de rejet des eaux usées, en comparaison de ceux qui étaient approvisionnés par la compagnie des eaux Lambeth, dont la prise d'eau se trouvait en amont du point de rejet des eaux usées. Dans cette argumentation, John Snow accordait une telle attention aux facteurs de confusion que ses propos méritent d'être répétés.

> « Les tuyaux de chacune des compagnies sillonnent toutes les rues et quasiment toutes les cours et ruelles. Certaines maisons sont alimentées par une société et certaines par l'autre compagnie […] Étant donné qu'il n'y a par ailleurs aucune différence ni sur le plan des maisons ou des personnes qui sont approvisionnées par les deux compagnies des eaux, ni dans les conditions physiques environnantes, il est évident qu'il n'aurait pas été possible de concevoir une expérience capable de tester de manière plus approfondie l'effet de l'approvisionnement en eau sur la progression du choléra que celle-ci, qui s'est imposée naturellement aux yeux de l'observateur. »

Autrement dit, les conclusions crédibles de telles études ne découlent pas d'un tour de passe-passe mathématique ou d'une procédure technique, et les

économétriciens qui présentent leur domaine en ces termes n'ont pas compris que ce qui fait progresser leur science est de savoir distinguer entre les cas susceptibles de fournir des informations, et ceux dont on ne tirera manifestement rien. Au lieu de cela, les preuves crédibles sont souvent ancrées dans la compréhension scientifique sociale d'une situation, qui est soumise à une longue série de vérifications empiriques. C'est pour cela que bon nombre de ces études sont plus longues et plus détaillées que celles, plus simples, basées sur les calculs et présentées dans la littérature, qui utilisent l'assignation aléatoire.

Les preuves empiriques récentes issues de RCT de grande qualité

Il existe une quatrième catégorie : celle des RCT de grande qualité. En 2018, des résultats ont été obtenus dans le cadre de trois RCT menées dans trois pays en développement, chacune ayant été conçue pour mesurer l'effet d'un type d'intervention sanitaire en milieu rural sur la taille des enfants et sur d'autres critères. Deux de ces RCT font partie de l'essai multisites WASH-*Benefits* (désigné dans la suite de ce chapitre par l'abréviation WASH-B). L'une a été réalisée au Kenya (Null *et al.*, 2018), l'autre au Bangladesh (Luby *et al.*, 2018). La troisième rend compte de l'essai SHINE mené au Zimbabwe (Humphrey *et al.*, 2019). Malgré certaines différences entre les trois études, chacune d'entre elles s'est concentrée sur la santé et la nutrition des enfants en bas âge dans une zone rurale d'un pays en développement. Chacune visait à améliorer les conditions sanitaires du foyer ou du « complexe » dans lequel vivait un jeune enfant[2] et à promouvoir l'utilisation de latrines et les comportements en matière d'hygiène. Et chacune a comparé une intervention sanitaire à une intervention en matière de nutrition (ici au sens d'alimentation). Au final, aucune des trois n'a révélé un effet de leur traitement sanitaire sur la taille de l'enfant.

Au cours des mois qui ont suivi la publication des résultats de ces trois RCT de grande qualité, des débats ont émergé sur les conclusions que l'on pouvait en tirer (Cumming et Curtis, 2018). Avant même de se demander *pourquoi* les trois RCT avaient abouti à ces résultats, les chercheurs ont débattu de la question de savoir s'ils étaient surprenants ou attendus. Coffey et Spears (2018), par exemple, montrent que les enquêtes démographiques et de santé (EDS) menées dans les zones rurales des trois pays ne révèlent aucun lien entre la taille des enfants et les moyennes communautaires des variables testées, avec un contrôle des plus sommaires sur le statut socio-économique.

Arnold *et al.* (2018), en revanche, s'appuient sur le grand nombre d'études observationnelles mal structurées réalisées au niveau des foyers pour conforter l'argument selon lequel les études observationnelles présentes dans la littérature laissaient envisager que ces RCT donneraient des effets significatifs sur le plan

2. Dans le cadre du programme WASH-B, l'intervention sanitaire a permis d'équiper l'ensemble de la maisonnée (dont la taille varie entre 3 et 10 foyers au Bangladesh, et entre 1 et 4 foyers au Kenya) en toilettes améliorées. Dans le programme SHINE, l'intervention sanitaire a fourni des toilettes améliorées seulement au foyer de l'enfant.

statistique. Arnold *et al.* montrent en particulier que les enfants figurant dans la série de données de référence provenant des mêmes sites sont en moyenne plus grands dans les foyers qui disposent de latrines que dans ceux qui n'en possèdent pas. Sur la base de leurs comparaisons, ils concluent que le fait que leur RCT n'ait pas montré d'effet donne une raison de douter de toutes les études observationnelles en général, y compris des preuves empiriques issues de stratégies bien pensées, et même de celles obtenues en posant différentes questions. Il s'agit pour nous d'un argument fallacieux : il est *bien entendu* possible de construire une étude observationnelle non crédible, comme l'ont fait Arnold *et al.*, tout comme il est possible de mener une RCT bâclée qui ne donne pas d'informations. Mais la compréhension de chacun ne progressera que si l'on prend en compte les meilleures preuves empiriques que peut produire chacune des approches.

Dans la quête d'explications, il est apparu qu'il y avait davantage d'interprétations que d'études. Les RCT fournissent à elles seules peu d'indications sur la manière d'*utiliser* leurs estimations pour faire avancer la science et l'élaboration des politiques. Deaton et Cartwright (2018 : 18) méritent une citation complète :

> « Les résultats [des RCT] ne peuvent pas être utilisés pour aider à faire des prévisions au-delà de l'échantillon-test si l'on ne fait pas un travail de structuration plus poussé, si l'on ne dispose pas de plus amples informations préalables et si l'on n'a pas une idée de ce qui fait varier les effets du traitement d'un endroit à un autre ou d'une période à une autre. Il n'y a pas d'autre choix que de s'engager dans une certaine structure causale si nous voulons savoir comment utiliser les preuves des RCT hors du contexte d'origine. Une simple généralisation et une simple extrapolation ne suffisent pas. Cela est vrai pour toute étude, qu'elle soit expérimentale ou observationnelle. Mais les études observationnelles connaissent bien et s'appuient régulièrement sur le type d'hypothèses que les RCT prétendent éviter (sans y parvenir), de telle sorte que, si le but est d'utiliser des preuves empiriques, l'avantage de crédibilité éventuel que peuvent avoir les RCT en matière d'évaluation n'est plus valable. Et c'est parce que les RCT nous en disent si peu sur le « pourquoi » des résultats trouvés qu'elles sont désavantagées par rapport aux études qui utilisent un éventail plus large d'informations et de données préalables permettant de comprendre les mécanismes en jeu. »

Dans le cas des programmes SHINE et WASH-B, les principaux auteurs des études ont suggéré, dans une revue de la littérature (Pickering *et al.*, 2019), que leurs interventions spécifiques n'avaient pas montré d'effet parce que les foyers n'étaient pas assez souvent sollicités par des encouragements à modifier leur comportement. C'est possible. Une alternative est que les effets sanitaires proviennent d'externalités à l'échelle du village (Geruso et Spears, 2018) plutôt que de changements au niveau du foyer : le foyer d'un enfant ne constitue qu'une petite partie de l'environnement sanitaire global de l'enfant. Andrés *et al.* (2017) ont constaté que, dans les zones rurales de l'Inde, un enfant qui passe d'un

foyer sans installations sanitaires améliorées dans un village avec un faible taux d'adoption à un foyer disposant d'un système sanitaire amélioré dans un village avec un fort taux d'adoption peut voir la prévalence de la diarrhée diminuer de 47 %. Un quart de cette réduction peut être attribué au bénéfice direct de l'accès au système sanitaire, les trois quarts restants provenant de l'effet indirect des voisins qui utilisent des installations sanitaires améliorées. Le rapport entre la part du village disposant d'une installation sanitaire améliorée et la prévalence de la diarrhée semble non linéaire. Il n'y a pratiquement pas d'externalité dans les villages où le *take up* du système sanitaire est faible. Les meilleures données d'observation disponibles révèlent des effets émanant de tels changements liés à un contexte, une communauté, un voisinage ou un village, mais pas d'effets issus de comparaisons entre un foyer disposant d'une installation sanitaire plus salubre et un foyer voisin qui n'en dispose pas. Il est important de souligner que, dans les RCT mentionnées ci-dessus (SHINE et WASH-B), la dimension du groupe d'intervention était généralement inférieure à celle qui serait attendue si les externalités étaient pleinement prises en compte. Dans la plupart des RCT sur l'assainissement, la dimension du groupe d'intervention est un village. Dans l'essai SHINE, seul le foyer propre de l'enfant a été traité, alors que l'essai WASH-B a pris en compte uniquement des maisonnées représentant généralement deux foyers environ, en comptant celui de l'enfant. Ceux-ci introduisent des changements dans l'environnement de l'enfant beaucoup plus limités que ceux considérés comme étant significatifs dans les hypothèses formulées dans la littérature existante.

Une autre série de possibilités repose sur le fait que les effets des mesures sanitaires sont hétérogènes, et qu'il est donc inapproprié de les considérer comme des effets induits purement et simplement par l'« assainissement » (Cumming et Curtis, 2018). Certains chercheurs font par exemple valoir que c'est la défécation en plein air en particulier qui est nuisible à la santé des enfants, mais dans les zones rurales du Kenya et du Bangladesh où se sont déroulés les essais WASH-B, la défécation en plein air était déjà très faible au départ, à hauteur d'un seul chiffre de points de pourcentage.

Une autre hétérogénéité prise comme hypothèse dans la littérature concerne la densité de la population. En tant qu'indicateur de la taille des enfants et de la mortalité infantile, les taux de défécation en plein air locale interagissent avec la densité de population : la défécation en plein air est plus fortement associée à une taille inférieure des enfants dans les endroits où la densité de population est élevée (Hathi *et al.*, 2017). Dans les lieux où la densité de population est faible, les données observationnelles ne montrent en moyenne pas de rapport entre ces deux facteurs. Dans les endroits où la densité de population est aussi faible que dans les zones rurales du Zimbabwe et du Kenya, Hathi *et al.* montrent que les données observationnelles ne laissent entrevoir aucune relation entre la taille des enfants et le système sanitaire local, tout comme l'ont montré les études expérimentales.

La fig. 1 montre que les données observationnelles ne prédisent pas de lien entre le système sanitaire et la taille des enfants, quel que soit le contexte. La figure est

calculée à partir des données[3] de l'EDS du Zimbabwe rural, le site sur lequel a été réalisé l'essai SHINE. L'axe horizontal correspond au type de couverture sanitaire favorisé par l'expérience : système sanitaire « amélioré », c'est-à-dire ni défécation en plein air, ni utilisation de latrines simples non améliorées. En particulier, on fait la moyenne de la couverture sanitaire sur l'ensemble des foyers du village, une valeur de 0,4 indiquant ainsi que 40 % des foyers échantillonnés de manière aléatoire déclarent utiliser des installations sanitaires non améliorées. L'axe vertical correspond à la mesure du rapport taille-âge chez les enfants, qui est présente dans l'ensemble de ces publications. Chaque point représente un village rural. La relation parfaitement plate – qui correspond à la prévision de HATHI *et al.* (2017) dans un contexte de faible densité de population – démontre que, dans les zones rurales du Zimbabwe, les données observationnelles de qualité supérieure collectées au niveau de la population ne révèlent aucun lien entre le rapport taille-âge et cette mesure des conditions sanitaires. En revanche, les mêmes données issues de l'EDS montrent de fortes relations entre la défécation en plein air locale et la taille de l'enfant dans les zones densément peuplées de l'Inde et dans d'autres contextes. Bien entendu, la fig. 1 ne peut pas expliquer entièrement les résultats d'une étude, quelle qu'elle soit, mais elle contribue à réfuter les avis de certains défenseurs des RCT qui allèguent que les preuves récentes obtenues sont en contradiction avec les données observationnelles, ou que les données observationnelles prédisaient que l'essai SHINE pourrait montrer un effet. Corrélation ne signifie pas causalité, mais les données observationnelles pour le Zimbabwe ne révèlent aucune corrélation.

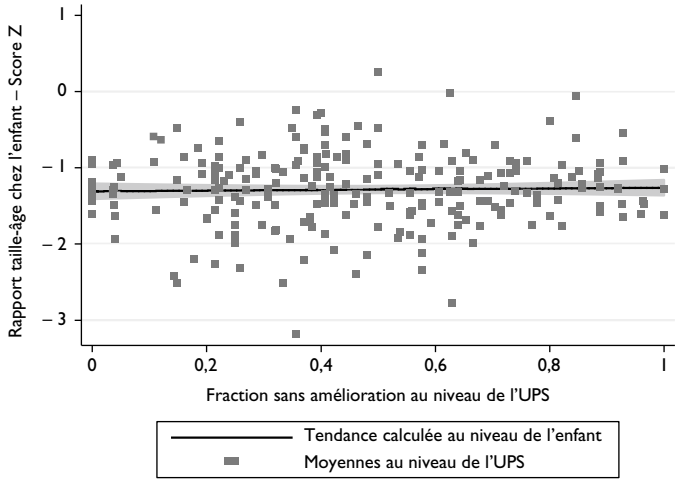


Figure 1

Lien entre les systèmes sanitaires améliorés et le rapport taille-âge chez les enfants au Zimbabwe.
Source : Dean Spears, Radu Ban et Oliver Cumming.

3. Les observations portent sur tous les enfants de moins de 60 mois vivant en zones rurales, pris en compte dans le recensement des naissances de l'EDS du Zimbabwe de 2015. Le rapport taille-âge est celui calculé par l'EDS, selon les normes de référence internationales de l'Organisation mondiale de la santé (OMS) de 2006. Chaque enfant est associé à un niveau d'hygiène moyen du foyer calculé pour tous les foyers pris en compte dans le recensement des foyers dans son unité primaire de sondage (UPS). Les points sont des moyennes d'UPS. La ligne, calculée au niveau des enfants, inclut un intervalle de confiance de 95 %.

Notre argument n'est pas de dire que nous savons laquelle de ces possibilités explique les résultats des programmes SHINE et WASH-B. Au contraire : personne ne peut le savoir avec certitude. Ce que nous voulons dire, c'est que cette situation fâcheuse n'est pas surprenante. En effet, les défis en matière d'assainissement en général ou pour le secteur WASH vont bien au-delà de cela. Le terme « assainissement » recouvre une multiplicité d'intervention. Il est également important de noter que, comme le soutient la publication de consensus (Cumming *et al.*, 2019), ces trois études ne remettent pas en cause les preuves empiriques (résultant d'études observationnelles à fort impact, telles que Cutler et Miller [2005]) indiquant que les améliorations à grande échelle des installations d'eau et d'assainissement ont contribué significativement à une meilleure santé des enfants dans la période de croissance des pays qui ont aujourd'hui un revenu élevé.

L'étalon-or ? Les défis des RCT dans le secteur de l'assainissement

Un cas paradigmatique de RCT dans le domaine de la santé est un essai clinique de médicament : dans un groupe de patients similaires, certains sont choisis de manière aléatoire pour prendre une pilule, tandis que d'autres sont sélectionnés de la même façon pour recevoir un placebo. Avaler une pilule suffit pour que le médicament soit totalement administré. Les pilules reçues par le groupe de traitement sont chimiquement identiques à tout autre cas d'administration de ce médicament. Cette méthode est peut-être bien l'« étalon-or » pour la recherche médicale, bien que Deaton et Cartwright (2018) contestent également cette interprétation. Dans cette partie, nous montrons que la métaphore n'est pas aussi prometteuse en ce qui concerne les systèmes sanitaires dans les pays ruraux en développement.

L'hétérogénéité des paramètres : des RCT différentes dans le secteur WASH devraient donner des réponses différentes

Certains des articles les plus cités en médecine et en épidémiologie sont des méta-analyses associant de multiples études. Reconnaissables à leurs « graphiques en forêt » avec des intervalles de confiance superposés, celles-ci combinent les estimations issues de nombreuses études individuelles pour calculer l'ampleur de l'effet et l'intervalle de confiance. Le calcul d'une estimation globale avec un graphique en forêt présente l'avantage majeur de réduire l'incertitude liée à l'erreur d'échantillonnage. Même si l'échantillon de chaque étude individuelle est petit et que son intervalle de confiance est donc important, l'estimation globale peut présenter un petit intervalle de confiance. En d'autres termes, plusieurs estimations imprécises de la même quantité peuvent, lorsqu'elles sont associées, fournir une estimation précise de cette quantité.

Là encore, le cas paradigmatique est une série d'essais cliniques d'un seul médicament chimiquement identique sur des populations comparables. Dans ce cas, les différentes RCT sont en effet des estimations multiples de la même quantité. Les études WASH, qu'elles soient ou non randomisées, ne sont quasiment jamais aussi uniformes dans les questions qui sont posées. L'hétérogénéité des paramètres ne constitue pas en elle-même une critique des RCT[4]. Toutefois, *dans la pratique*, les RCT auront du mal à traiter de manière appropriée l'hétérogénéité des effets si les études sont limitées en nombre, en raison de leurs coûts élevés et de leur mise en œuvre chronophage. En outre, une RCT réalisable sur les conditions sanitaires en milieu rural doit être limitée à un ou à quelques contextes, contrairement aux études observationnelles qui peuvent être représentatives de l'hétérogénéité existant au sein de populations entières, voire de groupes de pays. Nous passons ci-après en revue quelques exemples d'hétérogénéité.

Type de WASH

L'un des Objectifs de développement durable (ODD) est de mettre fin à la défécation en plein air ; un autre est de fournir un accès sûr et abordable à l'eau potable. Ces deux objectifs sont importants, mais les informations sur la manière d'atteindre l'un d'entre eux ou sur les effets que pourrait avoir sa réalisation ne sont pas directement pertinentes pour l'autre. Plus généralement, le secteur WASH intègre les réseaux d'égouts urbains et les systèmes de traitement, l'hygiène des mains (avec ou sans savon, avec ou sans agents antimicrobiens), le traitement de l'eau (à la source, à domicile, avec traitement physique, purification chimique ou désinfection solaire, sélectif ou non sélectif), les toilettes (latrines non améliorées, latrines améliorées, etc.), et bien d'autres choses encore.

Type d'assainissement

La différenciation entre la défécation en plein air, l'utilisation de latrines non améliorées et l'utilisation de latrines améliorées constitue un sous-ensemble de la différenciation au sein du secteur WASH, mais nous la mettons en évidence parce qu'elle est très souvent négligée. Les études WASH-B, par exemple, ont été menées dans des contextes qui, avant le début de l'expérience, ne comportaient déjà pas (ou peu) de défécation en plein air. Elles ne peuvent donc pas fournir d'informations directes sur les avantages de la réduction de la défécation en plein air dans les contextes où celle-ci reste courante, comme dans l'Inde rurale.

Lorsque les études WASH-B ont été publiées, *The Hindu* (grand journal indien) a fait paraître un article intitulé « Le lien entre assainissement et retard de croissance remis en question » (Pulla, 2018). Dans son paragraphe d'introduction, l'article faisait référence à la défécation en plein air, une question politique très médiatisée dans l'Inde d'aujourd'hui. Mais l'article n'a jamais expliqué que les études WASH-B, sur lesquelles il se focalisait, ne portaient pas du tout sur la défécation en plein air, qu'il n'y avait plus (ou peu) de pratiques de défécation

4. Par « hétérogénéité », nous entendons des différences dans les conditions locales, et non des différences dans les réponses individuelles aux interventions.

en plein air au Bangladesh, ou que dans l'Inde rurale, en revanche, la plupart des foyers défèquaient alors en plein air.

Certes, il serait peu probable que d'éminents chercheurs confondent eux-mêmes différentes catégories de problématiques sanitaires. Les journalistes, quant à eux, seraient tout à fait susceptibles de donner de fausses informations sur des preuves empiriques randomisées et non randomisées. Peu de chercheurs, quelle que soit leur méthodologie, sont enthousiastes à l'idée de corriger un journaliste au moment toujours excitant où une étude attire enfin l'attention du public. Cependant, comme les RCT sont rares et coûteuses (en temps et en argent), elles font naître une furieuse envie de leur donner publiquement une importance pouvant aller bien au-delà du sujet qu'elles traitent effectivement.

Effet communautaire ou propre au foyer

Il peut arriver que les enfants d'un foyer soient affectés par des germes introduits dans l'environnement par le comportement sanitaire d'autres foyers. Si tel est le cas, le fait d'éliminer la défécation en plein air uniquement dans le propre foyer de l'enfant, par exemple, est susceptible de ne pas modifier significativement son environnement sanitaire. Il peut ainsi y avoir des effets liés à la fois aux conditions sanitaires du foyer propre d'un enfant et à la couverture sanitaire au niveau de la communauté.

Densité de population, contexte urbain/rural et autres environnements

La même différence dans l'exposition aux conditions sanitaires peut induire une différence plus ou moins grande dans l'exposition aux agents pathogènes fécaux, dans des contextes où les gens vivent plus ou moins proches ou éloignés les uns des autres. Comme nous l'avons vu, certains chercheurs comme Hathi *et al.* (2017) ont constaté que la défécation en plein air interagit avec la densité de population pour prévoir la mortalité infantile et la taille des enfants.

Contexte sanitaire et comportements des populations en matière de santé

Les interventions WASH peuvent avoir des effets différents selon l'état de santé de départ de la population. Les interventions de déparasitage, par exemple, ne seront guère utiles dans un contexte où il n'y a pas d'infections par les vers. Par ailleurs, il a été démontré que l'eau potable interagissait avec l'allaitement maternel.

Ce qui est peut-être le plus pertinent dans notre contexte est le fait que beaucoup d'études et de rapports sur les politiques conceptualisent la taille des enfants en termes de *retard de croissance*, plutôt que de la considérer comme une variable continue. Le retard de croissance est une mesure dichotomique de la taille qui divise les enfants en deux catégories : les enfants beaucoup trop petits et les autres. Le fait de « dichotomiser » la taille de cette façon réduit la puissance statistique d'une étude à mettre un effet en évidence, même s'il y en a bien un (Spears *et al.*, 2013). Cela peut en outre engendrer une hétérogénéité dans les

paramètres : une intervention sanitaire pourrait augmenter le rapport taille-âge moyen dans la même mesure dans deux populations différentes, mais avoir des effets très différents sur le retard de croissance selon que l'enfant moyen est proche ou loin de la limite du retard de croissance.

Facteurs culturels et sociaux

Une même intervention sanitaire peut avoir des conséquences différentes si elle est mise en œuvre dans des sociétés différentes. Cette possibilité a été particulièrement bien documentée en Asie du Sud, où la caste (Lamba et Spears, 2013) et la religion (Ghosh *et al.*, 2014) ont toutes les deux été considérées comme susceptibles de prédire et, plus généralement, de façonner les pratiques de défécation en plein air. Les musulmans de l'Inde sont par exemple en moyenne plus pauvres et plus défavorisés en matière de services publics que les hindous, mais ils ont moins tendance à déféquer en plein air. Du fait de la ségrégation résidentielle, il est plus probable que les enfants musulmans vivent à proximité d'autres enfants musulmans (et vice versa), la conséquence étant que les enfants musulmans ont plus de chances que les enfants hindous de survivre à leur première année de vie. Avant que l'on ne dispose de l'explication basée sur les conditions sanitaires au niveau communautaire, les démographes appelaient cette énigme le « paradoxe de la mortalité musulmane » (Geruso et Spears, 2018). D'autres facteurs entrent probablement aussi en ligne de compte. Nous observons que les différences mises en lumière dans cette liste ont une pertinence scientifique et politique indéniable. L'utilisation des RCT dans le secteur WASH pour la mise en œuvre de politiques basées sur des preuves empiriques a au moins deux implications complexes. La première concerne le rapport coûts-bénéfices d'un investissement dans une évaluation randomisée plutôt que dans d'autres modes d'administration de la preuve. Une RCT, si elle est mise en œuvre avec succès, aura peut-être l'avantage de la clarté des conclusions sur les causes et les effets (voir toutefois Deaton et Cartwright [2018] avant d'en être certain). L'ensemble des cas auxquels s'adressent les RCT peut toutefois être plus restreint que dans le cas paradigmatique d'un essai clinique de médicament en raison de toutes ces sources d'hétérogénéité. L'autre défi est que l'intégration d'une telle hétérogénéité dans une étude est exactement le genre de chose que peut faire une étude observationnelle de grande qualité, en exploitant de vastes échantillons de données démographiques et des variations entre les contextes. En effet, les contextes culturels et environnementaux, la densité de population et la répartition de base des tailles constituent autant de facteurs qui ne peuvent pas être randomisés, et qui doivent être interprétés par le biais d'études observationnelles ou des compléments observationnels à une étude d'intervention.

Les erreurs de type 3, les premiers stades faibles et les traitements qui ne traitent pas

Les études observationnelles tentent de tirer des informations des variations qui existent déjà dans le monde. L'avantage est que l'on dispose de cas de comparaison couvrant l'ensemble des bonnes et mauvaises conditions sanitaires,

souvent au sein d'un même pays. En d'autres termes, il existe déjà de grandes différences dans l'exposition aux conditions sanitaires. Certains défis peuvent bien entendu subsister si les différences d'exposition aux conditions WASH sont corrélées aux variations dans d'autres dimensions favorables ou défavorables, et c'est pour cela que les meilleures études observationnelles recherchent des cas particuliers qui peuvent être instructifs et les étudient de près. Les évaluations randomisées d'interventions tentent de *générer* des variations dont on peut tirer des informations, et cela peut constituer un défi si ces variations sont difficiles à générer.

Prenons un exemple extrême. Dans la politique de développement actuelle, quasiment tout le monde s'accorde à dire que l'élimination rapide de la défécation en plein air est une grande priorité. Mais imaginez que ce ne soit pas le cas, et que les décideurs essaient de définir si l'élimination de la défécation en plein air doit être considérée comme une priorité sur la base de ses conséquences sur la santé. Imaginez également que les décideurs s'accordent à dire que les seules preuves admissibles seraient celles émanant de RCT. Supposons enfin que (parce que l'élimination de la défécation en plein air n'était pas déjà une priorité politique) personne ne connaissait encore les techniques susceptibles de réduire avec succès la défécation en plein air. Dans une telle situation, il se pourrait que la défécation en plein air ait des conséquences très importantes sur la santé dans certains contextes, mais qu'aucune preuve admissible ne puisse démontrer les avantages pour la santé de manière suffisamment convaincante pour justifier d'investir dans la façon d'introduire des variations dans la défécation en plein air.

Le secteur WASH ne se trouve heureusement pas tout à fait dans une situation aussi paradoxale. Toutefois, dans le cas de la défécation en plein air, en particulier, il a été difficile de générer des variations dans l'exposition à cette pratique, surtout et précisément dans les contextes où l'effet pourrait être le plus important. Du fait de l'histoire du système des castes et de l'importance persistante de l'intouchabilité, l'Inde rurale s'est avérée extrêmement résistante au changement en matière de défécation en plein air (Coffey et Spears, 2017). Clasen *et al.* (2014) ont mené une étude minutieuse et de grande qualité sur la défécation en plein air dans la région rurale d'Odisha, un État pauvre de l'Inde. Malheureusement, la conclusion des auteurs est que l'intervention n'a pas entraîné de changement majeur dans le comportement de défécation en plein air, et n'a donc pas introduit de variations suffisantes dans l'exposition à la défécation en plein air pour être instructive. Cependant, si la défécation en plein air s'avère plus nocive pour la santé des enfants dans les régions à forte densité de population, l'incapacité à produire des preuves de type *RCT à partir des cas étudiés dans l'Inde rurale* conduira à ce que la littérature néglige ce qui pourrait être l'effet le plus important, et ce dans le contexte même où la défécation en plein air reste la plus courante. De façon apparemment paradoxale, les endroits où les problèmes persistent ne sont pas nécessairement ceux où les décideurs politiques ne disposent pas des outils permettant de les résoudre. Il ne serait cependant pas surprenant que cela se révèle souvent être le cas.

Ainsi, une étude d'intervention ne peut pas produire d'informations si elle *tente* de générer une différence, mais n'y parvient pas. L'étude de CLASEN *et al.* (2014), dans laquelle les participants ne se soumettent pas au traitement et ne modifient pas leur comportement, est un exemple de ce problème. Dans ce cas, les chercheurs peuvent tirer une information utile sur ce qui ne fait pas changer les comportements, mais ils n'apprendront rien concernant les effets sur la santé. Le problème peut survenir sous une autre forme, même lorsque l'étude est mise en œuvre exactement comme prévu et que le comportement change strictement comme espéré, mais que l'étude n'est pas conçue au bon niveau de traitement. Prenons par exemple les essais SHINE et WASH-B, qui n'ont traité qu'un ou deux foyers dans le village des enfants étudiés. Si le facteur pertinent pour déterminer l'exposition est la situation sanitaire au niveau de la communauté, ces interventions n'auraient alors pas généré de grandes différences dans l'exposition, même si elles avaient été parfaitement mises en œuvre.

En collaborant avec d'autres disciplines pour écrire ce chapitre, nous avons appris que ce phénomène était connu sous différents noms. La littérature médicale et épidémiologique utilise parfois l'expression « erreur de type 3 » (en analogie avec les erreurs inférentielles familières dites de type 1 et de type 2) pour décrire une intervention qui ne s'est pas déroulée de manière effective. Dans ce cas, l'erreur consisterait à tirer une conclusion sur ce que *serait* l'effet de l'intervention proposée à l'origine si elle avait lieu. La littérature économique définit le même phénomène comme étant un « premier stade faible ». Dans un certain nombre d'études, le traitement randomisé n'est qu'un outil, ou un « instrument », comme on dit dans ce domaine, pour connaître l'effet d'une variable de premier stade sur un résultat de second stade. Un traitement randomisé peut par exemple tenter de générer des différences dans l'exposition à la défécation en plein air (premier stade) afin de connaître l'effet de la défécation en plein air sur la taille de l'enfant (second stade). Si le premier stade n'est pas suffisamment solide ou statistiquement clair (il existe des évaluations formelles pour définir un premier stade faible), alors l'étude ne sera pas informative sur l'effet du second stade considéré.

Sur le plan conceptuel, l'argument est simple : si une étude d'intervention ne génère pas ou ne peut pas générer une grande différence dans l'exposition à un type de condition sanitaire, alors elle ne donnera pas beaucoup d'informations (pour ou contre) sur l'effet de cette condition sanitaire. Les statisticiens compliquent cet argument simple en introduisant la notion de « calculs de puissance » : la *puissance* d'une étude est la probabilité de détecter un effet, s'il existe effectivement. Une étude comprenant un petit échantillon et un premier stade faible aura probablement une faible puissance. Cela induirait un large intervalle de confiance pour l'estimation finale de l'effet. Parler d'un « large intervalle de confiance » n'est qu'une façon élégante de dire que l'étude n'a pas permis d'apprendre grand-chose : on ne peut pas exclure qu'elle ait des effets importants, ou pas d'effets du tout. Ainsi, par exemple, lorsque des variables instrumentales sont utilisées pour produire un intervalle de confiance pour l'effet de la défécation en plein air locale sur la taille des enfants, et sur la base

des données issues de la RCT menée dans l'Odisha (anciennement Orissa) par Clasen *et al.* (2014), on obtient un intervalle de confiance très large. L'intervalle de confiance inclut la possibilité que la défécation en plein air ait des effets très importants sur la taille de l'enfant, la possibilité qu'il n'y ait pas d'effet du tout, et même la possibilité d'effets pervers dans lesquels la défécation en plein air améliore la situation des enfants. Un large intervalle de confiance est simplement une façon pour les statisticiens de dire qu'ils n'ont pas appris grand-chose.

Nous pouvons estimer la puissance d'expérimentations menées dans le domaine sanitaire qui n'ont pas eu lieu en faisant des hypothèses plausibles et en exploitant les données démographiques existantes. Geruso et Spears (2018) calculent par exemple la dimension de l'échantillon qui serait nécessaire pour estimer l'effet des externalités de la défécation en plein air sur la mortalité infantile. La mortalité infantile est une variable de résultat au moins aussi importante que la taille de l'enfant, mais elle est plus difficile à étudier dans un échantillon de dimension modérée parce qu'elle est « dichotomisée », et parce que les décès de nourrissons sont trop courants, mais rares sur le plan statistique. Une étude dimensionnée de manière adéquate nécessiterait un échantillon très large. Geruso et Spears calculent qu'une telle intervention, dans leur contexte indien, coûterait environ 90 millions de dollars, sans compter les frais de collecte, de gestion et d'analyse des données, en supposant des effets de premier stade optimistes sur les comportements sanitaires.

Dans la fig. 2, nous nous livrons à des calculs similaires pour des expériences hypothétiques afin d'étudier l'effet d'externalités de la défécation en plein air sur la taille des enfants. Arnold *et al.* (2011) décrivent en détail une stratégie de base pour estimer la puissance statistique d'une étude à l'aide de simulations randomisées. La figure utilise l'EDS 2005-2006 de l'Inde. Les simulations qui sous-tendent la figure supposent qu'il existe un effet important, réel, uniforme, constant et linéaire de l'exposition à la défécation en plein air : des écarts-types du rapport taille-âge de 0,5, associés linéairement au passage de 100 % à 0 % de la défécation en plein air dans le village des enfants étudiés. Les hypothèses concernant le « premier stade » de l'étude varient de deux façons. Premièrement, on suppose que l'intervention produit l'une des quatre intensités d'effet sur le comportement de défécation en plein air, avec une réduction de la défécation en plein air dans les foyers traités de 30 %, 50 %, 70 % et 90 %, représentées par les quatre lignes dans chaque graphique. Deuxièmement, dans le graphique (a), tous les foyers des villages d'intervention ont été traités avec l'intervention, mais dans le graphique (b), seuls les foyers ayant un enfant dont la taille a été mesurée ont été traités. Ainsi, dans le graphique (b), seule une petite fraction des foyers est supposée être traitée (bien que cette fraction soit toujours plus importante que dans les essais SHINE et WASH-B).

Dans chaque graphique, l'axe horizontal représente le nombre supposé de villages. L'axe commence à 400 villages, ce qui constituerait une expérience de grande ampleur selon les normes en vigueur dans la littérature. L'axe vertical représente la fraction des RCT simulées dans laquelle un effet de dimension statistiquement significative a été détecté. Dans tous les cas, les données sont

construites en supposant un effet important et constant, de sorte que l'axe vertical indique la probabilité de constater un effet qui est réellement présent.

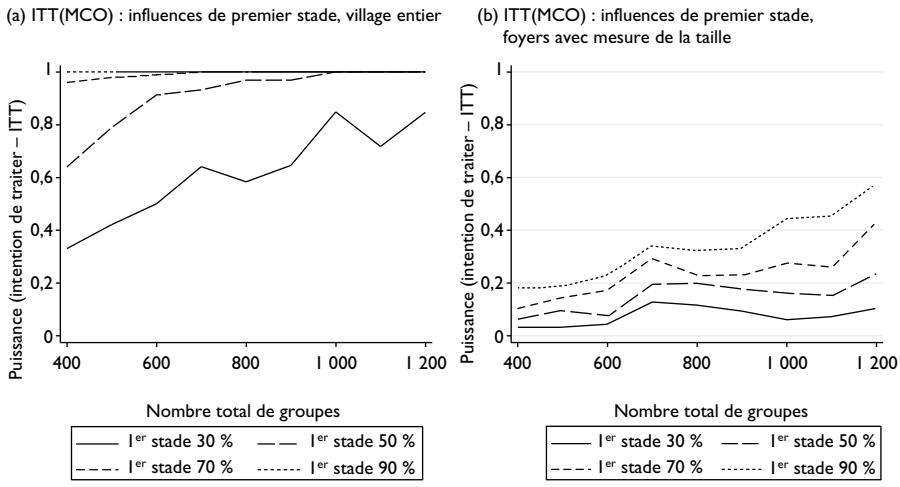


Figure 2

Simulations de la puissance d'expériences sanitaires hypothétiques en Inde rurale (méthode Monte-Carlo) selon différentes hypothèses concernant l'effet de premier stade sur la défécation en plein air dans les villages.

Source : Dean Spears, Radu Ban et Oliver Cumming.

Les chercheurs retiennent parfois 80 % comme niveau raisonnable de puissance pour qu'une expérience puisse détecter un effet réellement présent. Dans le graphique (a), où tous les foyers d'un village sont traités, les études peuvent atteindre ce seuil si la dimension de l'échantillon est suffisamment grande, si l'effet de premier stade est suffisamment important, ou les deux. Le graphique (b) montre que même des échantillons de très grande dimension – beaucoup plus grande que ce qui pourrait vraisemblablement être financé –, avec des effets de premier stade importants, ne révèlent pas une puissance suffisante si seuls les foyers avec enfants sont traités par l'intervention sanitaire. Bien sûr, ces simulations sont le reflet des hypothèses utilisées : dans le cas présent, l'hypothèse centrale est l'effet de grande ampleur sur les conditions sanitaires au niveau du village, plutôt qu'au niveau des foyers. Sur la base de ces hypothèses, la fig. 2 montre que la puissance d'une RCT, même très coûteuse, nécessaire pour détecter un effet dépend absolument de l'intensité et de la nature du premier stade.

Les sujets importants qui ne seront pas randomisés

Dans ce chapitre, nous avons souligné qu'il existe des études de qualité supérieure et inférieure pour chaque méthodologie, et que les RCT et les études observationnelles comportent une série de défis propres qui ne se recoupent pas. Nous souhaitons répondre ici à l'argument extrême selon lequel il convient de

se méfier de *toutes* les preuves empiriques non issues de RCT. Voici quelques exemples de questions qui sont importantes pour les politiques d'assainissement en milieux urbains et ruraux dans les pays en développement, et qui ne pourront jamais trouver de réponse seulement avec une RCT.

– Quels sont les effets sur la santé de l'utilisation accrue de systèmes sanitaires urbains « gérés de manière sûre[5] » (soit par la modernisation des égouts et des installations de traitement des eaux usées des villes, soit par l'amélioration de la gestion des matières fécales sans égouts) ?

– Plus largement, quelles sont les conséquences d'un système sanitaire inclusif à l'échelle de la ville, qui ne peut pas être seulement partiel ?

– En Inde, les musulmans sont-ils plus susceptibles d'utiliser des latrines que les hindous ? Pourquoi ?

– L'effet de la défécation en plein air est-il différent dans les villes ou dans les lieux à forte densité de population ?

– Quel est l'effet de la défécation en plein air sur la mortalité infantile, une variable dépendante dichotomique rare qui est très importante, mais impossible à étudier de manière réaliste avec une RCT pour des raisons de puissance statistique ?

– Quels sont les effets à long terme d'un bon système sanitaire au fil des générations, afin que les enfants naissent de mères dont le propre milieu utérin ne souffre pas de pertes nutritives liées à des agents pathogènes fécaux ?

– Quelle est l'ampleur de la défécation en plein air dans les foyers qui possèdent des latrines, même si ces latrines sont utilisées par certaines personnes ? Où ? Pourquoi ?

– Le manque d'accès à la défécation en plein air constitue-t-il un problème particulier pour les pauvres, les femmes ou les personnes âgées ? Les coûts d'un système sanitaire inadapté sont-ils plus élevés pour ces groupes de personnes ?

Les RCT ne peuvent pas, à elles seules, répondre à ces questions, soit parce qu'elles ne portent en définitive pas sur la relation de cause à effet, soit parce qu'elles concernent un facteur (comme la culture ou le lieu) que l'on ne peut pas faire varier expérimentalement. En outre, et en particulier dans les zones urbaines (par opposition aux zones rurales), l'assainissement ne relève pas seulement d'une intervention au niveau des foyers, consistant à construire et à utiliser des latrines, mais plutôt d'un système de gestion des flux de déchets humains (au moyen d'égouts souterrains ou par leur vidange, leur transport et leur traitement *via* des installations non reliées en réseau). Un tel système inclut le comportement des ménages, mais aussi l'infrastructure et la réglementation, et ces deux derniers éléments ne se prêtent pas à une manipulation expérimentale[6].

5. Comme défini dans l'indicateur ODD 6.2.

6. Bien que cela dépasse le cadre de ce chapitre, nous remarquons que l'efficacité du développement des infrastructures (qu'il s'agisse de systèmes sanitaires, de routes, de services publics ou de soins de santé) est essentielle pour réduire la pauvreté. Pourtant, elle n'est intrinsèquement pas du ressort des RCT.

Les RCT peuvent aider à répondre à un certain nombre de ces questions, mais pas toutes seules. Dans le domaine du développement, le « mouvement pour des politiques basées sur des preuves empiriques » est quasiment allé de pair avec l'offensive des RCT dans l'élaboration des politiques. Mais de nombreux décideurs manquent même de preuves empiriques capables de décrire où et dans quelles populations se situent les défis, par exemple pour connaître le niveau de défécation en plein air rencontré dans divers districts de l'Inde. D'autres questions concernent forcément les interactions. Si la densité de population ou les pratiques culturelles interagissent effectivement avec un traitement en influant sur son efficacité, alors un décideur (comme la Banque mondiale ou la fondation Gates) doit comprendre cette interaction, ce qui exige nécessairement de comprendre des variables observationnelles non randomisées.

Même la série standard de questions sur la santé peut nécessiter un temps de réponse plus long que ce qui est généralement admis. Il est par exemple possible qu'il existe des voies intergénérationnelles par lesquelles la malnutrition nette est transmise : une mère en mauvaise santé dans son enfance peut grandir et présenter un milieu utérin plus petit, avec des impacts sur la croissance de l'enfant. Par ailleurs, l'exposition aux maladies à l'âge adulte, si elle réduit la masse corporelle de la mère avant la grossesse ou sa prise de poids pendant la grossesse, pourrait rendre le corps de la mère moins apte à nourrir l'enfant *in utero* et pendant l'allaitement. Ce constat demeurerait identique si une étude d'intervention hypothétique éliminait totalement les agents pathogènes fécaux immédiatement avant la conception de l'enfant. Les études observationnelles, en revanche, peuvent tirer des informations sur les variations d'équilibre à long terme selon l'exposition aux conditions sanitaires.

Les questions négligées qui pourraient progresser grâce aux RCT (ou à des études d'intervention ne relevant pas de RCT)

Nous constatons enfin que certaines questions de santé publique pourraient progresser en réalisant un plus grand nombre et une plus grande variété de RCT, notamment celles qui portent sur la manière de modifier les comportements des individus ou d'autres agents économiques importants. Dans le domaine de l'économie du développement, BLATTMAN (2008) appelle ces types de RCT « évaluations d'impact 2.0 ». Il compare la typologie 2.0 des RCT, qui se concentrent sur *comment* et *pourquoi* certaines interventions fonctionnent, à la typologie 1.0, qui vise à déterminer *si* les interventions fonctionnent. Dans le même ordre d'idées, DUFLO (2017) fait remarquer que les économistes (du développement) doivent « adopter l'état d'esprit d'un plombier », en concentrant leurs recherches sur la façon d'améliorer les problématiques de prestation de services dites « du dernier kilomètre ». Les études d'intervention[7] ciblant

7. Nous utilisons le terme « étude d'intervention » pour désigner une étude dans laquelle une intervention est conçue ou modifiée spécifiquement dans le but d'informer sur son [suite p. suiv.]

les changements de comportements pourraient être plus limitées, plus rapides et moins coûteuses que les études sur la santé qui s'étendent sur plusieurs années. Elles pourraient être répétées par un processus « d'essais et d'erreurs », comme le recommandent Pritchett *et al.* (2013) dans leur description de « l'apprentissage dans les projets de développement ». Dans de nombreux cas, il serait opportun de randomiser ces essais, même si le plus important est d'enregistrer avec soin les leçons tirées de la mise en œuvre. Ce n'est pas sans une certaine ironie que nous recommandons aux chercheurs du secteur WASH d'adopter l'état d'esprit du plombier lors de la conception des études d'intervention.

Pour rendre cette recommandation plus concrète, considérons les ODD dans le domaine sanitaire, qui visent à mettre fin à la défécation en plein air et à augmenter la couverture de systèmes sanitaires gérés de manière sûre. L'un des axes recommandés pour ces études d'intervention serait par conséquent de réduire les comportements de défécation en plein air dans les lieux où cela s'est avéré difficile, comme dans les zones rurales de l'Inde (Rosenboom et Ban, 2017). La série d'études organisées par l'International Initiative for Impact Evaluation (3ie), la fondation Bill et Melinda Gates et le Research Institute for Compassionate Economics (RICE), destinée à tester des stratégies de promotion de l'utilisation des latrines dans l'Inde rurale, a constitué un effort notable vers la réalisation de cet objectif. Les processus qui sous-tendent ces études et leurs résultats brossent un tableau peu réjouissant de la façon dont même les RCT « 2.0 » les mieux conçues, axées sur des sujets de recherche bien définis et guidés par les politiques, se débattent parfois dans les difficultés pour produire des preuves pertinentes pour les politiques. Après un processus itératif de recherche formative, certaines études pilotes prometteuses ont été sélectionnées pour des études d'intervention complètes. Les quatre études d'intervention complètes ont été randomisées. Mais, même ce processus bien élaboré s'est avéré très long et ses résultats difficiles à interpréter. Bien que le projet ait débuté en 2015 dans le but de fournir des informations au programme indien Swachh Bharat Mission (SBM), ses résultats n'ont été disponibles qu'au second semestre 2019, et avaient donc peu de chance de renseigner la mission avant sa conclusion en octobre 2019. En outre, les résultats des quatre études (Friedrich *et al.*, 2019 ; Chauhan *et al.*, 2019 ; Visawanathan *et al.*, 2019 ; Caruso *et al.*, 2019) restent difficiles à interpréter. Elles ont montré des augmentations significatives de l'utilisation des latrines (et des réductions de la défécation en plein air) à la fois dans le groupe de traitement et dans le groupe de contrôle, probablement en raison de la forte ampleur du programme national SBM, et/ou du fait du « biais de courtoisie » accru provoqué par la forte exposition aux messages encourageant l'utilisation des latrines.

efficacité. Toutes les études d'intervention ne sont pas des RCT. Les études d'intervention, selon notre interprétation, diffèrent des études observationnelles. Ces dernières se servent des variations qui surviennent naturellement pour déterminer l'efficacité des interventions.

Il serait également recommandé que ces études d'intervention se concentrent sur la bonne gestion de l'assainissement[8], en particulier dans les zones urbaines où l'action de nombreux agents aux comportements multiples est requise pour minimiser le rejet d'excréments non traités dans l'environnement immédiat. HOUDE *et al.* (2017) étudient par exemple une intervention au niveau de la fourniture de services qui consiste à augmenter la concurrence entre les chauffeurs de camions vide-fosses afin de réduire le coût d'une bonne vidange des fosses septiques. Il convient de noter que la réalisation de l'une de ces études d'intervention coûte entre 10 et 30 fois moins cher que les études WASH-B ou SHINE.

Conclusion : la bonne utilisation de bonnes preuves empiriques est la seule norme applicable

Dans le secteur WASH pour les pays en développement, il n'existe pas d'autre étalon-or qu'une recherche minutieuse et réfléchie. Dans la pratique, cela requiert la collaboration de chercheurs d'horizons différents possédant des compétences différentes. Si le nombre de mots dans les revues spécialisées dans les sciences sociales empiriques est généralement plus élevé que les 4 500 demandés par le *Lancet Global Health*, nous espérons que cela ne reste pas un obstacle important à la collaboration. Il est également nécessaire de porter un jugement sur les meilleures et les pires modes d'administration de la preuve, et sur les contextes auxquels les preuves empiriques sont susceptibles de s'appliquer. Deaton et Cartwright ne seraient pas surpris par l'expérience du secteur WASH : « donner un statut spécial aux RCT est injustifié. La méthode la plus susceptible de produire une bonne inférence causale dépend de ce que nous essayons de découvrir, et de ce qui est déjà connu. Lorsque l'on dispose de peu de connaissances préalables, aucune méthode n'est susceptible de produire des conclusions bien étayées. »

Pour certains chercheurs, la solution réside dans un plus grand nombre de RCT. D'après les articles publiés des essais SHINE et WASH-B, les enquêtes de terrain en amont des essais ont commencé respectivement en 2011 et 2012. La réflexion, la planification, la recherche de financements et l'embauche de

8. Les études entrant dans le cadre de ce volet thématique de 3ie, qui se sont achevées en août 2019, révèlent un autre défi pratique dans l'utilisation des RCT pour déterminer l'efficacité d'une intervention dans le contexte d'un programme très vaste et important sur le plan politique. La mise en œuvre par le gouvernement du programme Swachh Bharat Mission a pris une telle ampleur que des réductions de la défécation en plein air ont été observées à la fois dans les populations de traitement et dans les populations de contrôle (probablement parce que l'efficacité du programme gouvernemental était supérieure à celle des interventions spécifiques ciblées dans la RCT, et/ou parce que la défécation en plein air autodéclarée était extrêmement sensible aux biais, les individus concernés étant conscients qu'ils devaient déclarer utiliser des latrines).

personnel ont certainement débuté plusieurs années auparavant. Les résultats ont été rendus publics en 2018. Ces études sont des réalisations impressionnantes et complexes. Leur réalisation s'est étendue sur près d'une décennie. Pour le meilleur ou pour le pire, ces études ont été menées à une période où les politiques et les pratiques en matière d'assainissement évoluaient rapidement dans les pays en développement. Selon les statistiques de l'United Nations International Children's Emergency Fund (Unicef) et de l'OMS, la proportion de personnes dans le monde disposant d'un système sanitaire géré de manière sûre augmente de près d'un point de pourcentage chaque année. Les ODD prévoient l'élimination totale de la défécation en plein air d'ici 2030. Une autre série de trois essais aussi complexes que SHINE et WASH-B pourrait occuper la majeure partie de la décennie qui reste avant cette échéance.

Le secteur WASH dispose heureusement de nombreuses « connaissances préalables ». Si les trois RCT récentes ont suscité de nombreuses discussions, quasiment personne parmi ceux qui croyaient auparavant que l'exposition aux agents pathogènes fécaux nuisait au développement des enfants n'a changé d'avis. Peut-être que ces études nous disent de ne pas investir d'argent dans la transformation de latrines non améliorées en latrines améliorées. Peut-être soulignent-elles l'importance de la densité de population, des nombreuses visites nécessaires pour entreprendre des changements de comportements, ou des externalités au-delà du seul foyer de l'enfant. Personne ne le sait pour l'instant, et les RCT n'ont pas tranché ces questions. Cette conversation va se poursuivre, et elle continuera à s'appuyer sur diverses sources permettant d'apporter des preuves empiriques, comme cela devrait toujours être le cas.

Chapitre 7

Les expérimentations aléatoires en microfinance

Miracle ou mirage ?

Florent Bédécarrats, Isabelle Guérin et François Roubaud

Introduction

L'essor du microcrédit et la généralisation des évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) ont marqué deux avancées majeures des politiques de développement de réduction de la pauvreté au cours des dernières décennies (Cling *et al.*, 2003). L'essor du microcrédit a eu lieu dans les années 1990. Son point culminant date du début des années 2000, avec le lancement par l'ONU de l'année internationale du microcrédit (2005) et l'attribution du prix Nobel de la paix à Mohammad Yunus et à la Grameen Bank, qu'il a fondée. Les RCT ont connu un succès retentissant dix ans plus tard, lorsqu'elles sont devenues la méthode de référence par excellence pour les évaluations d'impact et quand, en 2019, un autre prix Nobel (d'économie cette fois-ci) a été décerné à Esther Duflo, Abijit Banerjee et Michael Kremer, chefs de file du mouvement RCT. De fait, ces deux avancées sont étroitement corrélées : le microcrédit a été l'un des thèmes phares, un sujet emblématique, des interventions évaluées par les expérimentations randomisées dans le domaine du développement.

Ce chapitre présente un examen détaillé des RCT sur le microcrédit en s'appuyant sur un large éventail d'outils analytiques employés en statistique, en économie politique, en sociologie et en anthropologie du développement. Il se concentre principalement sur le numéro spécial (ci-après dénommé « numéro spécial ») publié en 2015 dans une grande revue d'économie, *American Economic Journal: Applied Economics* (*AEJ:AE*). Ce numéro spécial réunit six RCT sur le microcrédit et les articles sont préfacés par une introduction générale (ci-après appelée « introduction générale ») qui tire des conclusions communes. Le numéro spécial

a rencontré un large écho dans les cercles académiques et professionnels, tant et si bien qu'il a tendance à être considéré comme la conclusion définitive sur les impacts (limités) du microcrédit. Mais l'est-il vraiment ?

Nous abordons ce numéro spécial sous deux angles : (1) *top-down*, avec un test sur un cas précis (microcrédit) illustrant les critiques générales faites contre les RCT, en particulier celles formulées par les auteurs dans un précédent article (Bédécarrats *et al.*, 2019b) et (2) *bottom-up*, avec une étude de la mise en œuvre des RCT sur le terrain. Nous prenons comme point de départ une réplication de l'une des six RCT abordées dans le numéro spécial : la RCT conduite dans le Maroc rural (Bédécarrats *et al.*, 2019a ; 2019b), qui joue un rôle central dans l'« économie » du numéro spécial. Nous élargissons ensuite la perspective du cas marocain pour adopter un angle plus général en identifiant les invariants présents dans d'autres RCT et en déterminant les particularités de chacune d'entre elles. De façon plus générale, la principale question que nous posons dans ce chapitre est la suivante : « Quels enseignements peut-on tirer des RCT sur le microcrédit et comment peut-on expliquer leur succès mondial alors qu'elles souffrent d'un manque évident de rigueur ? »

Le reste de ce chapitre est organisé comme suit. Après avoir résumé les principales caractéristiques des six expérimentations, la deuxième partie présente leurs principaux résultats et replace le numéro spécial dans le contexte général du poids et du rôle du microcrédit dans l'industrie des RCT. La troisième partie adopte un point de vue comparatif pour identifier les principales critiques techniques qui peuvent être formulées contre ce corpus d'expérimentations, tant en termes de validité interne que de validité externe, ainsi que les problèmes éthiques soulevés. Au-delà de la méthode et des résultats quantitatifs, la quatrième partie analyse les interprétations proposées par les auteurs (notamment dans l'introduction générale) et leur théorie sous-jacente du changement. Dans la conclusion, nous proposons une interprétation du hiatus évoqué ci-dessus – un succès de grande envergure malgré des lacunes importantes – et tirons des enseignements plus généraux de notre travail.

RCT et microcrédit : un produit phare en désuétude ?

Le microcrédit figure parmi les principaux services fournis par la microfinance, l'un des secteurs les plus fréquemment évalués par les RCT. Cette importance est illustrée par la base de données en ligne des RCT gérées par le Abdul Latif Jameel Poverty Action Lab (J-PAL) – un centre de recherche mondial qui préconise cette méthode pour la réduction de la pauvreté et qui, de fait, est le principal fournisseur et promoteur des RCT. En 2010, cette base de données comptait 233 RCT, dont 32 % classées sous le label « microfinance » (Bédécarrats,

2012). Depuis lors, le J-PAL a revu son système de classement en élargissant ses catégories et compte actuellement 287 RCT classées en « finance » sur un total de 978 RCT[1]. De fait, la finance est le secteur d'intérêt par excellence du J-PAL, devant l'éducation (233 RCT) et « l'économie politique et la gouvernance » (216 RCT). Bien que la microfinance ne soit qu'un sous-ensemble des RCT « finance », le J-PAL produit un grand nombre d'évaluations d'impact sur le sujet. Les RCT sur la microfinance et l'industrie générale des RCT ont véritablement pris leur envol au milieu des années 2000 (Bédécarrats *et al.*, 2019b ; Ravallion, chap. 1, ce volume). Depuis, le nombre de RCT consacrées à la microfinance a fortement régressé, alors que les RCT en général ont poursuivi leur expansion (fig. 1). Notons que le comptage des RCT conduites à travers le monde est un défi. Nos estimations sont fondées sur la base de données d'évaluations d'impact en ligne de l'International Initiative for Impact Evaluation (3ie), complétée par Bédécarrats (2012) et sur la base de données d'évaluation en ligne du J-PAL[2]. Comme l'indique la fig. 1a, l'impact de la microfinance constitue de longue date un sujet controversé, source de nombreuses évaluations d'impact non expérimentales. Alors que les méthodes expérimentales produisent des preuves empiriques quantitatives théoriquement plus solides, les études non expérimentales offrent pléthore de preuves pertinentes. Une forte hausse des évaluations expérimentales a également été observée, coïncidant avec une diminution nette des évaluations non expérimentales, même si ces tendances peuvent être marginalement exagérées par l'omission des études les plus récentes dans les référentiels que nous avons utilisés. La fig. 1b montre par ailleurs que la microfinance a constitué un thème prédominant du mouvement *randomista*[3] jusqu'en 2013, mais que cet intérêt s'est depuis affaibli. Le déclin observé dans la seconde moitié des années 2010, après le pic de la première moitié de cette décennie est quelque peu intrigant : est-il dû à une inversion de tendance ou est-ce parce qu'il n'y a plus grand-chose à dire sur ce sujet, désormais trop rebattu ? C'est un point que nous traiterons plus en détail par la suite.

1. Source : site internet de The Abdul Lateef Jameel Poverty Action Lab : www.povertyactionlab.org/evaluations.

2. La base de données d'évaluations d'impact en ligne du 3ie constitue le principal catalogue de résultats d'évaluations d'impact conduites sur des interventions de développement (https://www.3ieimpact.org/evidence-hub/impact-evaluation-repository, consulté le 13/10/2019). Le 3ie a tendance à omettre des évaluations non expérimentales et son travail d'inventaire semble avoir perdu de sa vigueur ces dernières années, les références se faisant plus rares à partir de 2015. Nous avons complété les données du 3ie par les évaluations d'impact recensées dans Bédécarrats (2012) et par les références incluses dans la base de données d'évaluation du J-PAL. Les références ont été appariées afin d'éviter de compter deux fois les mêmes évaluations. La fig. 1b est basée sur les références recensées dans la base de données d'évaluation en ligne du J-PAL (https://www.povertyactionlab.org/evaluations, consulté 18/10/2019). Le terme « finance » dans la clé est le label assigné par le J-PAL à l'évaluation enregistrée. Les auteurs ont attribué le label « microfinance » après examen des résumés de toutes les évaluations enregistrées en « finance » sur le site du J-PAL. Les dates de la fig. 1b indiquent l'année d'achèvement de l'expérimentation, tandis que les dates de fig. 1a indiquent l'année de publication de ses résultats.

3. Nous appelons « *randomistas* » les promoteurs de RCT qui sont convaincus que les RCT constituent le seul moyen d'évaluer rigoureusement l'impact dans l'évaluation et sont en tout point supérieures aux autres méthodologies.

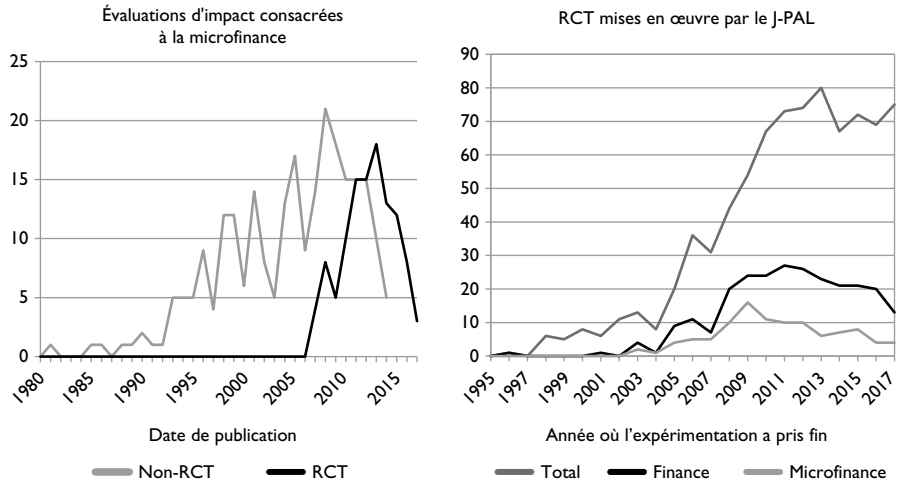


Figure 1
RCT sur la microfinance.

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base du référentiel d'évaluations 3ie (2019), du référentiel d'évaluations J-PAL et de Bédécarrats (2012) pour le panel 1a ; et sur la base du référentiel d'évaluations en ligne J-PAL pour le panel 1b.

C'est à l'apogée des RCT consacrées à la microfinance qu'un numéro spécial a été publié en 2015 dans l'*American Economic Journal: Applied Economics*, présentant six RCT sur le microcrédit (Banerjee *et al.*, 2015c). Pour les chefs de file du mouvement RCT, ce numéro spécial est jugé comme une contribution décisive, permettant de clore un débat (Ogden, 2017) qui a longtemps fait rage tant dans les cercles académiques que parmi les bailleurs et les décideurs politiques. Ce numéro spécial a rapidement bénéficié d'une couverture massive, comme en témoignent les 3 607 citations de ses articles parus dans d'autres publications scientifiques[4]. Soucieux d'en tirer des leçons opérationnelles à l'intention des décideurs politiques, le J-PAL et l'Innovations for Poverty Action (IPA) en ont tiré une note de synthèse suggérant des conclusions générales et universelles, valables pour le microcrédit dans le monde entier (J-PAL et IPA, 2015). Certains chercheurs ont même spéculé que ce pourrait être là le « dernier mot sur le microcrédit » (Sandefur, 2015).

Si l'on examine de plus près l'impact académique du numéro spécial de l'*AEJ:AE*, le résultat est saisissant. Selon Google Scholar (consulté le 13/10/2019), l'introduction générale a été citée pas moins de 527 fois. C'est là une belle performance, bien que loin derrière l'article de Banerjee *et al.* (2015b) sur le programme de microcrédit Spandana en Inde (1 813 citations). Les cinq autres articles ont également obtenu de très bons scores : 320 citations pour Angelucci, Karlan et Zinman (2015) sur le Compartamos Banco au Mexique, 298 pour Crépon *et al.* (2015) sur *Al Amana* dans le Maroc rural, 225 pour Attanasio *et al.* (2015) sur

4. https://scholar.google.fr/scholar?as_q=microcredit+OR+microfinance&as_epq=&as_oq=&as_eq=&as_occt=any&as_sauthors=&as_publication=american+economic+journal+applied&as_ylo=2015&as_yhi=2015&hl=en&as_sdt=0%2C5.

la Mongolie, 214 pour AUGSBURG *et al.* (2015) sur la Bosnie-Herzégovine et 210 pour TAROZZI *et al.* (2015) sur l'Éthiopie. À titre de comparaison, l'article de PITT et KHANDKER (1998), cité par ROODMAN et MORDUCH (2014) comme étant l'article empirique sur le microcrédit le plus cité de tous les temps, affiche 1 956 citations, plus de vingt ans après sa publication.

Outre ces citations directes, l'impact du numéro spécial s'illustre par des citations en cascade (à l'instar de tout article), mais aussi par des examens systématiques ou des méta-analyses qui, pour la plupart, trouvent dans le numéro spécial l'essentiel de leur corpus de preuves (BRODY *et al.*, 2015 ; BUERA *et al.*, 2015 ; DEMIRGUC-KUNT *et al.*, 2017 ; CHERNOZHUKOV *et al.*, 2018 ; MEAGER, 2019). Il convient de mentionner en particulier l'article publié dans la prestigieuse revue *Science* en 2015 (BANERJEE *et al.*, 2015a, cité 484 fois)[5]. Cet article traite en détail du numéro spécial, en mettant en exergue les mérites comparatifs d'une approche distincte (programmes dits de « graduation »)[6].

Enfin, les résultats du numéro spécial ont largement dépassé les cercles académiques pour pénétrer le monde des praticiens de la microfinance (J-PAL et IPA, 2015). Le Consultative Group to Assist the Poor (CGAP), qui joue un rôle de premier plan dans la diffusion des bonnes pratiques dans le secteur du microcrédit, l'a commenté avant même sa publication (CULL *et al.*, 2014). Pour de nombreux praticiens (que l'un d'entre nous rencontre régulièrement dans des conférences et sur le terrain), les résultats du numéro spécial relèvent désormais du sens commun.

En définitive, qu'elles soient jugées sur le nombre de RCT ou sur la diffusion de leurs résultats, les évaluations d'impact dans le domaine de la microfinance et, en particulier, du microcrédit, apparaissent comme le produit phare du mouvement *randomista*, et le numéro spécial comme le prototype par excellence de ce produit phare.

Gros plan sur la conception du numéro spécial de l'*AEJ:AE*

Le numéro spécial comporte six articles sur six RCT sur le microcrédit, menées par six équipes affiliées au J-PAL dans six pays différents (Bosnie-Herzégovine, Éthiopie, Inde, Mexique, Mongolie et Maroc) à peu près au même moment (de 2006 à 2012). Ces articles sont précédés d'une introduction générale qui tire des enseignements généraux de cette expérience collective. Le numéro spécial puise sa force dans un processus d'harmonisation organisé en aval par le journal en vue de sa publication[7]. Un plan d'analyse commun a été mis au point pour faciliter les comparaisons. Dans la mesure du possible, l'impact du microcrédit a été estimé

5. Ce n'est pas la première fois que *Science* ouvre ses colonnes aux RCT sur le microcrédit (KARLAN et ZINMAN, 2011).

6. Les programmes dits de « graduation » consistent en un ensemble séquencé d'interventions ciblant les ultra-pauvres et extrêmement pauvres qui sont enfermés dans les situations de pauvreté les plus difficiles.

7. « Grâce aux efforts déployés par les six équipes de recherche et la rédactrice en chef, Esther Duflo, pour rendre les articles facilement comparables, il a été relativement aisé de tirer des enseignements des six études » (BANERJEE *et al.*, 2015c : 2).

selon la même méthodologie économétrique afin d'obtenir un ensemble de résultats communs, eux-mêmes tous calculés de la même manière. C'était la première fois qu'un tel effort de mutualisation était déployé à cette échelle. Dans l'optique d'une montée en généralisation, cette mutualisation constitue un avantage décisif.

Ainsi, le numéro spécial n'est pas seulement déterminant en termes de résultats, il marque aussi un changement de « bonnes pratiques » de la part des promoteurs des RCT. Le numéro cherche en effet à remédier à un certain nombre de limites. Pour la première fois, le numéro dans son ensemble, et l'introduction générale en particulier, apporte des éléments de réponse à cinq types de critiques récurrentes du mouvement pro-RCT (Bédécarrats *et al.*, 2019b) : un modèle théorique est développé en réponse à la critique accusant les RCT d'empirisme agnostique ; une analyse coût-bénéfice est proposée pour répondre à la question de l'efficacité, afin de dépasser le simple impact causal ; les questions du taux d'adhésion, de la précision des estimateurs et de l'hétérogénéité des traitements sont reconnues et discutées ; la diversité contextuelle est abordée par une série de paramètres, de produits et d'institutions couverts par les six articles, permettant aux éditeurs du numéro spécial d'affirmer que leur échantillon est « assez représentatif de l'industrie/du mouvement du microcrédit dans le monde » (Banerjee *et al.*, 2015c : 2) ; et, enfin, le numéro spécial se propose de mettre à disposition les bases de données originales en réponse aux requêtes de réplication et afin de faciliter les méta-analyses.

Décrivons brièvement les six RCT. En dépit d'un processus d'harmonisation en amont (traitement et analyse des données), les expériences diffèrent grandement dans leurs protocoles. Les types de produits de microcrédit, les institutions de microfinance (IMF), l'unité des procédures de randomisation, etc. varient d'une RCT à l'autre. Les auteurs interprètent cette diversité en partant de l'hypothèse que la similitude des résultats dans ce large éventail d'environnements est une garantie de leur robustesse et met donc en évidence les propriétés génériques des impacts du microcrédit. En d'autres mots, ils entendent répondre à la critique récurrente déplorant le manque de validité externe des RCT.

L'introduction générale présente en détail les principales caractéristiques des six RCT, résumées dans le tabl. 1. Les IMF varient en taille, certaines ont une vocation commerciale et d'autres non. Toutes sortes de produits sont proposés : des prêts collectifs et individuels, des remboursements hebdomadaires et mensuels, un taux d'intérêt annuel allant de 12 à 110 % (en moyenne) et un montant de prêt (moyen) représentant 6 à 118 % du revenu mensuel. La moitié des programmes de microcrédit ciblent des femmes. Au niveau géographique, une RCT est exclusivement urbaine (Inde), trois sont exclusivement rurales (Éthiopie, Mongolie et Maroc) et les deux dernières couvrent les deux types de zones. Il convient de noter que, dans tous les cas, les critères d'éligibilité des clients sont *ad hoc* : ils dépendent à la fois des règles internes de chaque IMF et des paramètres de chaque RCT. Ce faisant, les populations cibles sont extrêmement spécifiques (voire uniques), ce qui restreint les possibilités d'inférence et d'extrapolation à des populations plus larges ; nous reviendrons sur ce point dans la troisième partie.

Tableau 1
Principales caractéristiques des six RCT.

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Taux annuel effectif global (TAEG)	22 %	12 %	24 %	110 %	27 %	14 %
Garantie	Individuelle	Solidaire	Solidaire	Solidaire	Les deux	Solidaire
Prêt/revenu du ménage moyen	9 %	118 %	22 %	6 %	43 %	21 %
Sexe des clients potentiels	Les deux	Les deux	Féminin	Féminin	Féminin	Les deux
Éligibilité au prêt (entre autres)	Garantie solide, capacité de remboursement, solvabilité…	Statut de pauvreté, *business plan*…	Entre 18 et 59 ans, justificatif de propriété du logement…	Entre 18 et 60 ans, carte d'identité en cours de validité, justificatif de domicile…	Actifs < 869 $ Bénéfices < 174 $/mois	Entre 18 et 70 ans, carte d'identité, activité agricole autre qu'élevage…
Couverture géographique (urbaine/rurale)	Les deux	Rurale	Urbaine	Les deux	Rurale	Rurale
Couverture géographique (en nombre de régions ou villes)	14 (à l'échelle nationale)	2 (Ouest)	1 (ville)	4 (NC Sonora)	5 (Nord)	11 (à l'échelle nationale)
Unité de randomisation	Individuelle	Association	Quartier	Quartier et village	Village	Village
Unité d'échantillonnage finale	Candidat risqué et jugé insolvable…	Ménages aléatoires	Ménage avec >= 1 femme >= 3 ans dans la zone…	Avoir une entreprise ou souhaiter en avoir une…	Être intéressé par l'obtention d'un prêt…	Ménage considéré comme emprunteur potentiel…
Taille de l'échantillon (enquête finale)	995	6 263	6 862	16 560	964	5 551

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base de Banerjee *et al.* (2015c, tabl. 1 et 2).

Exploration détaillée du numéro spécial de l'*AEJ:AE* : principaux résultats

L'introduction générale tire sept grands enseignements de l'exercice. En premier lieu, le faible niveau d'adhésion est une constante dans toutes les études, sauf en Bosnie-Herzégovine, ce qui amène à la conclusion que le microcrédit ne saurait faire figure de panacée pour sortir les populations de la pauvreté. Une conséquence préjudiciable de cette faible adhésion est qu'elle pose un problème de puissance statistique et remet en cause la stratégie d'identification des RCT. Toutefois, l'introduction générale prétend que les RCT du Maroc, de l'Inde et du Mexique apportent de nouveaux éléments pour remédier à ces lacunes (prévision du taux d'adhésion et stratégie d'échantillonnage). Deuxièmement, et dans le prolongement du point précédent, il est particulièrement difficile de prévoir le taux d'adhésion, et aucune étude n'est parvenue à le faire de façon satisfaisante. Troisièmement, et c'est probablement la principale conclusion, l'accès au microcrédit n'est pas transformateur, que ce soit au niveau des performances des micro-entreprises ou des conditions de vie des ménages – y compris concernant le « bien-être social » et « l'autonomisation » des femmes –, du moins en moyenne. Le seul résultat robuste pour la consommation est une baisse des dépenses dites « facultatives », à savoir, selon les auteurs, les « biens de tentation, loisirs/divertissements/célébrations » (Banerjee *et al.*, 2015c : 13). Quatrièmement, seul l'investissement des entreprises est augmenté par le microcrédit, ce qui montre que ce dernier stimule les intentions des micro-entrepreneurs visant à développer leur entreprise. Cinquièmement, d'autres effets modestes, mais potentiellement importants sont relevés : la liberté de choix en particulier. Sixièmement, bien que le microcrédit ne soit pas transformateur, il n'a pas non plus d'effets catastrophiques, ce qui place les partisans et les adversaires du microcrédit sur un pied d'égalité. Enfin, le septième enseignement concerne la présomption d'hétérogénéité de l'impact du microcrédit, qui pourrait être positif (voire transformateur) pour certains (la couche supérieure) et négatif pour d'autres. Cela nous ramène à la question de la puissance statistique, de la taille d'échantillon requise pour estimer correctement les impacts, et de la représentativité des populations ciblées. Le tabl. 2, basé sur l'introduction générale et le J-PAL et IPA (2015), résume les résultats obtenus par les six RCT au regard des principaux effets suivis.

En conclusion, le numéro spécial est considéré par beaucoup, à commencer par les auteurs eux-mêmes (Ogden, 2017), comme la synthèse la plus complète qui soit sur l'impact du microcrédit. Ses conclusions générales n'ont guère été remises en cause depuis sa publication en 2015 (pour les exceptions, voir Wydick, 2016 ; Dahal et Fiala, 2020). D'une certaine manière, il fige l'état des connaissances acquises sur les impacts causaux du microcrédit et sur son rôle dans le développement et l'éradication de la pauvreté. Pour les éditeurs de l'*AEJ:AE*, et les articles ultérieurs traitant des six RCT, le numéro spécial va encore bien plus loin. Il est salué pour avoir repoussé les frontières de la connaissance scientifique, tant en ce qui concerne le microcrédit que la méthode RCT. Trois papiers, postérieurs au numéro spécial et traitant directement du même ensemble de RCT, illustrent bien cet état de fait. L'article de Meager (2019), publié lui aussi dans l'*AEJ:AE*, confirme que le numéro spécial reste

incontournable. Cet article reprend les six RCT du numéro spécial (plus une RCT aux Philippines ; KARLAN et ZINMAN, 2011) pour réestimer l'impact général sur les principales variables et répondre à la question de la validité externe à l'appui d'une méthode innovante (une analyse hiérarchique bayésienne). Il y a ensuite CHERNOZHUKOV *et al.* (2018), qui appliquent une méthode de double apprentissage automatique pour étudier l'hétérogénéité de cet ensemble de données. Un troisième exemple est celui de BANERJEE *et al.* (2019a), publié au moment où ce chapitre était en cours de rédaction. Leur article s'appuie sur une troisième vague d'enquête pour la RCT indienne de Spandana. Tout en répondant à certaines des critiques formulées contre les RCT (en abordant le traitement hétérogène, en allongeant la durée et en développant un modèle théorique), l'article puise largement dans le numéro spécial, qui est ici présenté comme la somme des connaissances réunies à ce jour sur le microcrédit. Cet article n'est certainement pas le dernier. Dans la même veine, CRÉPON *et al.* (2015) annoncent dans la conclusion de leur article une troisième vague d'enquêtes pour la RCT marocaine afin d'évaluer l'impact à long terme du microcrédit[8].

Tableau 2
Principaux résultats des six RCT.

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Propriété d'entreprise	Positif	n.s.	n.s.	n. s.	Positif	n. s.
Revenus d'entreprise	n.s.	n.s.	n. s.	Positif	n. s.	Positif
Actifs d'entreprise	Positif	–	Positif	–	Positif	Positif
Investissements d'entreprise	n.s.	n. s.	Positif	Positif	–	Positif
Bénéfices d'entreprise	–	–	–	–	–	Positif
Revenus des ménages	n.s.	n.s.	n.s.	n.s.	n.s.	n. s.
Consommation des ménages	n. s.	Négatif	–	Négatif	Positif	–
Consommation des ménages en biens de tentation	Négatif	–	Négatif	Négatif	n. s.	Négatif
Bien-être social	n.s.	n.s.	n. s.	Positif	–	n. s.
Autonomisation des femmes	–	n. s.	–	Positif	–	–

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base de J-PAL et IPA (2015) ; BANERJEE *et al.*, (2015c).

Note : n.s. : non significatif à 10 % ; « – » : pas de données.

8. « Nous procédons actuellement à un suivi auprès des ménages, maintenant qu'un laps de temps beaucoup plus long s'est écoulé, pour vérifier si l'investissement dans les actifs commerciaux s'est révélé rentable sur le long terme » (CRÉPON *et al.*, 2015 : 148).

Validité et portée du numéro spécial : évaluation critique

Dans la littérature, les RCT sont évaluées sous deux angles principaux : la validité externe et la validité interne. La validité externe est essentielle lorsqu'il s'agit de transposer, d'informer et de concevoir des politiques publiques à une échelle plus large (nationale ou régionale) et de tester une théorie. La validité interne est généralement prise pour un fait acquis avec les RCT, et considérée comme leur principal point fort par rapport aux autres méthodes. Si cette caractéristique peut être vraie en théorie, les contraintes de mise en œuvre sur le terrain peuvent remettre en cause ces conditions idéales, un aspect jusqu'ici passé sous silence.

Validité interne

L'évaluation de la validité interne des RCT requiert des investigations sur l'élaboration, et le « bricolage », des RCT sur le terrain. Nous nous sommes livrés à cet exercice exigeant sur l'étude marocaine (Crépon *et al.*, 2015). Nous présentons ci-dessous les principaux résultats des deux articles complémentaires que nous avons produits à partir de cette étude (Bédécarrats *et al.*, 2019a ; 2019b).

Le cas emblématique de la RCT marocaine

De 2006 à 2010, une équipe de recherche du J-PAL a conduit une RCT dans le Maroc rural pour mesurer l'impact du microcrédit fourni par *Al Amana*, qui était à l'époque la principale IMF sur un marché marocain en pleine phase d'expansion.

Nous avons répliqué l'enquête de Crépon *et al.* et avons identifié un certain nombre de problèmes qui remettent en cause leurs conclusions (Bédécarrats *et al.*, 2019a). Les chercheurs ont utilisé des procédures et des seuils incohérents pour le *trimming* et leurs résultats sont lourdement tributaires de la manière dont les valeurs extrêmes ont été exclues. Crépon *et al.* (2015) font état d'un échantillon équilibré dans l'enquête initiale (EI), après avoir éliminé les valeurs extrêmes de 24 variables sur 459 observations (10,3 % de l'échantillon). Mais, dans l'enquête finale (EF), ils ont éliminé différemment 27 observations (0,5 % de l'échantillon) en les supprimant complètement. Le fait d'abaisser ce seuil dans l'enquête finale de seulement 0,2 % (en éliminant une douzaine d'observations en plus ou en moins) produit des résultats radicalement différents en termes de ventes, de dépenses, d'investissements et de bénéfices. Aucun autre seuil appliqué pour le *trimming* n'aurait donné des résultats conformes aux conclusions publiées et aucun autre article du même numéro spécial n'a eu recours à une méthode ou à un seuil similaire.

Nous avons constaté des déséquilibres substantiels et significatifs dans l'enquête initiale pour un certain nombre de variables importantes, dont les variables de résultats de la RCT. Dans ce contexte, nous avons estimé des « effets de traitement » peu plausibles sur certaines variables, par exemple le chef de ménage, le

sexe et la langue parlée. Nous avons en outre documenté de nombreuses erreurs de codage. Par exemple, l'évaluation des actifs agricoles dans l'enquête finale a omis deux types d'actifs (les tracteurs et les moissonneuses), qui se trouvent être les actifs les plus chers des ménages interrogés. Or, l'inclusion de tracteurs et de moissonneuses dans l'évaluation des actifs augmente la valeur moyenne des actifs agricoles par ménage de 470 % (de 1 377 à 5 111 dirhams). Les erreurs de codage identifiées ont altéré environ 80 % des observations.

Les incohérences des mesures du crédit méritent une attention particulière, car elles revêtent une importance primordiale pour caractériser le traitement évalué par cette expérimentation. Crépon *et al.* (2015) comparent les données administratives (celles de l'IMF) aux données de l'enquête, en préférant se baser sur le taux d'adhésion de 17 % des données administratives, contre les 11 % de l'enquête. Ils affirment que la population marocaine sous-estime les emprunts en raison de la honte religieuse qu'ils impliquent. Cependant, cette conclusion n'est pas plausible, car les incohérences entre les sources vont bien au-delà des différences de moyennes. Au total, 195 des 435 clients inclus dans les résultats ont affirmé n'avoir jamais emprunté à l'IMF. Or, cette explication de la « honte du crédit » pour ces ménages induirait une explication de « fierté du crédit » pour les 152 ménages ayant déclaré avoir obtenu un prêt de l'IMF sans figurer dans ses registres.

Par ailleurs, selon les données d'enquête recueillies sur l'échantillon du panel, l'accès au crédit est resté stable dans le groupe de traitement entre l'enquête initiale et l'enquête finale, alors qu'il a diminué dans le groupe de contrôle (la microfinance marocaine a traversé une crise majeure de 2008 à 2010). Nos résultats remettent en question le sens même de cette RCT : ce qui a été testé ne semble pas être l'impact de l'introduction du microcrédit dans des zones « vierges », mais plutôt le remplacement d'autres sources formelles par une source de microcrédit dans le groupe de traitement et un rationnement du crédit dans le groupe de contrôle.

Nous avons également trouvé des erreurs d'échantillonnage. Par exemple, la composition par sexe et par âge pour 20 % des ménages interrogés dans l'enquête initiale et réinterrogés dans l'enquête finale diffère à tel point qu'il est peu probable que ce soient les mêmes unités qui aient été réinterrogées. En outre, nous avons constaté que les caractéristiques de l'échantillon de Crépon *et al.* différaient sensiblement de celles de la population. La composition des ménages est passée de 5,17 à 6,13 membres entre l'enquête initiale et l'enquête finale. Or, le recensement national indique que les ménages ruraux marocains comptaient en moyenne 6,03 membres en 2004 et 5,35 membres en 2014. De telles divergences soulèvent des questions sur la représentativité de l'échantillon et compromettent donc la validité externe de cette étude.

Les auteurs ont produit une réponse à notre réplication, intitulée *Rejoinder*, réfutant la plupart des erreurs que nous avons documentées (Crépon *et al.*, 2019). Ils se sont référés à notre analyse, mais ne semblent pas avoir répliqué ou analysé de près son contenu statistique. Par ailleurs, le *Rejoinder* contient

de nombreuses erreurs et omissions factuelles. Nous avons publié un document passant en revue leurs principaux arguments en réponse à notre réplication (BÉDÉCARRATS *et al.*, 2019c). Nous avons constaté que toutes les erreurs de codage, de mesure et d'échantillonnage documentées dans notre réplication demeurent d'actualité.

Distorsion du protocole : ajustement du produit et de l'échantillonnage

Notre deuxième article a cherché à expliquer comment de telles incohérences pouvaient se produire, à l'appui d'une étude qualitative de terrain spécialement conçue pour compléter la RCT (MORVANT-ROUX *et al.*, 2014) et de différents documents et données publics et internes des principaux acteurs de la RCT (BÉDÉCARRATS *et al.*, 2019b). L'article décrit l'ensemble de la chaîne de production de l'étude, depuis l'échantillonnage, la collecte, la saisie et le recodage des données à la publication et à la diffusion des résultats en passant par les estimations et interprétations. Loin des conditions idéales de laboratoire[9], l'analyse de la mise en œuvre du protocole randomisé sur le terrain par les différents acteurs (chacun ayant ses propres motivations et contraintes) révèle un certain nombre de disparités par rapport au protocole théorique annoncé dans l'article publié.

L'un des principaux problèmes rencontrés durant l'étude a été le taux d'adhésion, bien inférieur aux prévisions initiales, ce qui a nécessité un certain nombre de mesures correctives. Le premier ajustement a consisté à modifier l'intervention (offre de microcrédit) en lançant de nouvelles campagnes d'information, en accordant des primes forfaitaires aux agents et en éliminant le quota minimum de femmes. L'adhésion est devenue une « obsession », tant pour l'équipe de recherche que pour les agents de crédit, qui ont eux-mêmes employé ce terme et se sont donné beaucoup de mal pour inciter les villageois à souscrire un microcrédit. Parmi les stratégies utilisées, il a été décidé de repousser les frontières habituelles des villages dans l'espoir de trouver davantage de clients[10]. Lorsque ces mesures se sont révélées insuffisantes, l'équipe a ajusté la méthode d'échantillonnage (modification des modèles prédictifs et ajout dans l'enquête finale de nouveaux ménages, censés présenter une plus grande propension à emprunter). Les villages affichant un taux d'adhésion nul ont été éliminés.

Mauvaise qualité des données et erreurs de mesure

La collecte et la saisie des données ont été sous-traitées à une société de conseil spécialisée dans l'ingénierie, mais sans aucune expérience des enquêtes statistiques. Afin de suivre la conception et la mise en œuvre de la RCT, le bailleur de fonds de la RCT (l'Agence française de développement – AFD) a nommé une

9. Les expérimentations de terrain telles que les RCT sont précisément conçues pour sortir du monde artificiel des laboratoires. Or, les *randomistas* ont trop souvent tendance à penser que le protocole peut être appliqué tel quel, comme en laboratoire, ce qui n'est pas le cas.

10. Un changement de produit pour l'intervention évaluée par la RCT pose également un problème de validité externe (car les conditions de l'expérimentation ne sont pas conformes à son fonctionnement dans le « monde réel ») (PETERS *et al.*, 2018).

équipe d'économistes et de spécialistes des enquêtes auprès des ménages. Lors de missions de terrain, cette équipe a très tôt constaté de graves dysfonctionnements dans la collecte des données. Ils ont observé des problèmes de traduction, car les enquêteurs ne parlaient pas le berbère, langue parlée par une grande partie de la population cible. Les enquêteurs ont donc eu largement recours à des traducteurs improvisés, parmi lesquels des dirigeants locaux, ce qui a induit des problèmes de compréhension et de biais dans les réponses (désirabilité sociale et méfiance à l'égard du gouvernement).

Un autre dysfonctionnement portait sur le nombre de répondants dans les ménages et les familles élargies, qui semblait ici encore relever de la pure improvisation, en fonction de la présence et de la disponibilité des personnes, mais aussi de leur capacité à se comprendre entre elles et à comprendre les enquêteurs. Ces observations expliquent probablement en partie les importantes divergences susmentionnées entre l'enquête initiale et l'enquête finale. Toutefois, l'ampleur des divergences suggère une autre explication : certains ménages n'étaient peut-être tout simplement pas les mêmes, comme le confirme notre réplication. L'absence d'adresse précise exige des techniques de suivi rigoureuses, qui ont peut-être été négligées. Faute de temps et de supervision, il est possible que certains enquêteurs se soient contentés d'interroger les ménages disponibles au moment de leur visite. L'équipe de l'AFD a formulé des recommandations pour améliorer la qualité des données collectées, en s'inquiétant des répercussions potentielles de ces lacunes sur les résultats de l'expérience. Elle a également dénoncé les problèmes de saisie des données. Bien que l'équipe du J-PAL ait répondu, contestant la gravité des problèmes et affirmant qu'ils ne remettaient pas en cause la validité interne de l'expérience, le comité de pilotage a décidé, lors de sa réunion suivante, que tous les questionnaires déjà saisis devaient être envoyés à l'Institut national français de la statistique (Insee) à Paris pour être à nouveau saisis.

Ces différents points n'ont pas été évoqués dans l'article publié et révèlent des lacunes dans la préparation, la mise en œuvre et le suivi du travail sur le terrain.

Au-delà de la RCT marocaine : évaluation générale

Il n'est pas possible d'analyser les cinq autres RCT du numéro spécial de manière aussi détaillée, aussi bien pour des raisons de temps que parce qu'il manque les données brutes pour deux d'entre elles (tabl. 3). Nous nous livrons donc à un exercice partiel, à savoir une lecture critique des principales caractéristiques techniques des procédures employées, en nous basant sur les informations disponibles dans les articles publiés. Le tabl. 4 résume les problèmes de validité interne tels qu'ils peuvent être évalués à partir des informations dont nous disposons. Presque aucun de ces problèmes n'est évoqué dans le numéro spécial, et encore moins dans l'introduction générale. Nous abordons tour à tour la question de l'erreur d'échantillonnage et la question de l'inexactitude de mesure.

Tableau 3
Validité externe, réserves reconnues et préoccupations éthiques.

	Bosnie-Herzégovine	**Éthiopie**	**Inde**	**Mexique**	**Mongolie**	**Maroc**
Population d'intérêt	Expansion de l'IMF	Expansion de l'IMF	Expansion de l'IFM (partielle)	Expansion de l'IMF	Expansion de l'IMF	Expansion de l'IFM (partielle)
Extrapolation à population plus générale ?	Non	Non	Non	Non	Non	Non
Discussion sur les limites potentielles (dans l'article) ?						
Hawthorne ou John Henry	Oui	Non	Non	Non	Non	Non
Équilibre général	Non	Non	Non	Non	Non	Oui
Comparaison avec des données d'OSN ?	Oui	Non	Non	Non	Oui	Non
Autres enquêtes/méthodes mises en œuvre ?	Non	Non	Non	Non	Enquêtes dans les villages, entretiens qualitatifs	Non
Si oui, utilisées ?	–	–	–	–	Non	–
Réserves explicites reconnues ?	**Oui** 1-Pas de validité externe 2-Sous-puissance statistique 3-Effets potentiels H & JH	**Oui** 1-Pas de validité externe 2-Sous-puissance statistique 3-Pas de panel = déséquilibre au niveau de l'EI, attrition sélective, effet hétérogène 4-Non-respect du plan expérimental 5-Pas de consommation 6-Erreurs de mesure	**Oui** 1-Sous-puissance statistique 2-EI non représentative 3-Attrition sélective et migration 4-Contamination 5-IdT représentative des « emprunteurs probables » uniquement	**Oui** 1-Pas de validité externe 2-Qualité des données 3-Pas d'EI 4-Périodes de traitement hétérogènes	**Oui** 1-Pas de validité externe 2-Sous-puissance statistique 3-Présence d'autres IMF 4-Attrition (déséquilibre possible) 5-Non robuste au TMH	**Oui** 1-Petits déséquilibres significatifs au niveau de l'EI

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Discussion sur les préoccupations éthiques						
Consentement éclairé à l'expérimentation	Non	Non	Non Non	Non	Non	Non
Analyse et suivi des risques	Non	Non	Non	Non	Non	Non
Équipoise	Non	Non		Non	Non	Non
Reproductibilité						
Données disponibles	Données brutes	Non	Données agrégées	Données agrégées	Données brutes	Données brutes
Code détaillé disponible	Oui	Non	Partiellement	Partiellement	Oui	Oui
Questionnaire d'enquête disponible sur le site d'AEJ	Oui	Non	Non	Non	Non	Oui

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base de Banerjee *et al.* (2015c).

Notes : IMF : institution de microfinance ; OSN : Office statistique national ; IdT : intention de traiter ; EI : enquête initiale ; EF : enquête finale ; H&JH : Hawthorne et John Henry ; TMH : test multi-hypothèses.

Tableau 4
Validité interne des six RCT.

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Population d'intérêt	**Clients potentiels** initialement refusés par l'IMF pour cause d'insolvabilité	**Ménages ruraux dans deux zones *ad hoc***	**Emprunteurs potentiels** (femmes vivant dans des bidonvilles depuis plus de trois ans avec une pièce d'identité valide) dans les zones d'expansion de l'IMF à Hyderabad	**Clients potentiels** (femmes possédant ou envisageant de créer une entreprise ou ayant l'intention d'emprunter) dans les zones d'expansion de l'IMF à Central Sonora, au Mexique	**Femmes pauvres :** (actifs < 869 $ et bénéfices < 174 $/mois) Inscrites pour obtenir un prêt	**Ménages à forte propension à l'emprunt dans les zones rurales d'extension de l'IMF**
Plan de sondage, randomisation						
Plan de sondage	Échantillon individuel à choix raisonné	– Stratifié (2 « zones ») – 3 degrés (unités administratives/village/ménage)	2 degrés (bidonvilles/ménages)	2 degrés (village/ménage)	– Stratifié (5 provinces) – 2 degrés (village/ménage) Mongolie du Nord	2 degrés (village/ménage)
Informations sur la sélection des zones	Non applicable	Oui	Oui	Oui	Non	Oui
Nombre de zones (T, C)		353 villages	104 (52, 52)	238 (120, 118)	25 (15, 10)	162 (81,81)
Zones abandonnées	Non applicable	Non	Oui (16 bidonvilles)	Oui (12 zones)	Non	Oui (non précisé)
Informations sur la sélection des personnes	Oui, non aléatoire	Oui, aléatoire	Oui, aléatoire	Oui, aléatoire	Oui, non aléatoire (30 premiers à s'inscrire)	Oui, aléatoire
Informations sur la randomisation (T *vs* C)	Oui (niveau individuel)	Oui (niveau du village)	Oui (niveau du bidonville)	Oui (niveau de la zone)	Oui (niveau du village)	Oui (niveau du village)

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Taille de l'échantillon (complet ; contrôle)	EI (1 196 ; 568) EF (994 ; 444)	EI (6 412 ; n.d.) EF (6 263 ; n.d.)	EI (2 800 ; 1 220) EF1 (6 863 ; 3 264) EF2 (6 142 ; 2 943)	EI (6 786 ; n.d.) EF (16 560 ; 8 298)	EI (710 ; 299) EF (610 ; 260)	EI (4 465 ; 2 266) EF (5 551 ; 2 810)
Taux d'attrition (EI->EF) : % total, % contrôle	Panel (17 % ; 22 %)	Pas de panel 2 groupes types	EI->EF : pas de panel EF1->EF2 (11 % ; 10 %)	Panel (37 % ; n.d.)	Panel (16 % ; 15 %)	Panel (8 % ; 7 %)
Respect du protocole expérimental	Oui	Non 22 % de zones mal allouées (12 % T non traités, 23 % C traités)	Non (16 zones supprimées ; EF non fiable)	Non (EI avortée)	Oui	Non (nouveaux ménages ajoutés à EF)
Tests d'équilibre dans l'enquête initiale						
Population incluse	Ménages du panel uniquement	Ménages du panel uniquement	Tous les ménages EI	Ménages du panel uniquement	Ménages du panel uniquement	Tous les ménages EI
Variables testées Inclusion des principaux résultats de l'étude	27 Oui	35 Oui	33 Oui	14 Non	48 Oui	43 Non
Déséquilibres importants signalés	Non	Non	Non	Oui	Oui	Oui
Trimming	Résultats avec et sans *trimming* de 1 % pour les contrôles de robustesse	Résultats avec et sans *trimming* 8 obs. pour les contrôles de robustesse	Non	Non	Non	EI : *trimming* des valeurs les plus élevées pour 10,3 % des obs. EF : *trimming* de 0,5 % des obs.
Qualité des données (discussion dans l'article)	Non	Oui, marginale (erreurs de mesure)	Oui, marginale (éventuelles erreurs de mémoire)	Oui, marginale (variables de résultats manquantes pour EF)	Non	Non (excepté pour l'adhésion) (données administratives *vs* enquête)

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base de Angelucci *et al.*, 2015 ; Attanasio *et al.*, 2015 ; Augsburg *et al.*, 2015 ; Banerjee *et al.*, 2015b ; Crépon *et al.*, 2015 ; Tarozzi *et al.*, 2015.

Note : EI : enquête initiale ; EF : enquête finale ; n.d. : non disponible ; obs. : observations ; T *vs* C : groupe de traitement par rapport au groupe de contrôle.

Concernant l'échantillonnage, il convient de noter que, de manière générale, les articles ne fournissent pas les éléments de base permettant de décrire et de qualifier avec précision les plans d'échantillonnage et de sélection adoptés (il existe pourtant des normes standardisées pour de telles descriptions, comme dans Statistique Canada, 2010 et Ardilly et Tillé, 2006). Les auteurs concentrent leurs analyses sur les questions de randomisation et d'inférence causale. Premièrement, la population de référence n'est jamais clairement établie. Dans la plupart des cas, elle correspond aux clients éligibles dans les zones d'expansion de l'IMF, quoique l'on ne sache pas comment ces dernières sont définies. Cela a des répercussions problématiques sur la validité externe des RCT (voir la partie « Validité externe » ci-après). Deuxièmement, les plans d'échantillonnage adoptés s'inscrivent dans la catégorie générale de l'échantillonnage aléatoire stratifié à plusieurs degrés, à l'exception des RCT en Bosnie-Herzégovine et en Mongolie. Aucun de ces deux cas ne fait en effet l'objet d'un échantillonnage aléatoire : en Mongolie, les trente premières femmes pauvres dans chaque village sélectionné à se dire intéressées par un prêt ont été retenues ; en Bosnie-Herzégovine, il a été demandé aux agents de crédit de sélectionner des clients potentiels qui n'étaient pas jugés éligibles selon les normes actuelles de l'IMF. Dans tous les cas, ces plans de sondage complexes, pour reprendre la terminologie statistique, soit ne permettent pas de calculer les intervalles de confiance associés à l'impact estimé (les deux cas mentionnés ci-dessus), soit nécessiteraient des calculs particulièrement complexes pour estimer la variance, calculs qui ne sont pas faits (sauf pour l'estimation des erreurs types robustes propres aux groupes). La conséquence directe de cette divergence est que les intervalles de confiance sont probablement sous-estimés et que les impacts réputés significatifs, déjà peu nombreux, ne devraient pas être statistiquement différents de zéro.

De plus, quatre des six RCT se sont écartées du protocole canonique de la méthode expérimentale : sélection aléatoire d'un groupe de traitement et d'un groupe de contrôle, enquête initiale avant traitement, puis suivi du panel basé sur une enquête finale après traitement. Dans le cas de l'Éthiopie, les enquêtes initiale et finale n'ont pas porté sur des panels, mais sur des enquêtes en coupe répétées (c'est-à-dire que différents individus ont été interrogés). Il apparaît donc impossible d'identifier dans l'enquête initiale les déséquilibres potentiels concernant la population pour laquelle l'impact est estimé dans l'enquête finale. Dans les cas du Mexique, du Maroc et de l'Inde, les enquêtes sur le terrain n'ont pas pu être menées comme prévu initialement et les risques pesant sur la réussite de l'expérience ont conduit à réajuster le protocole initial en cours de route. En Inde, l'enquête initiale n'a servi de panel de base pour aucune des deux enquêtes finales subséquentes[11], soulevant les mêmes problèmes que dans le cas de l'Éthiopie mentionné ci-dessus. Au Mexique, l'enquête initiale a avorté en raison de la mauvaise qualité des données recueillies : 73 % des

11. Banerjee *et al.* (2015b) ont conduit une première enquête finale en 2007 et 2008. Ils ont réinterrogé les ménages de la première enquête finale dans le cadre d'une deuxième enquête finale en 2009 et 2010.

ménages de l'enquête initiale n'ont pas été revus pour l'enquête finale et 89 % de l'échantillon final n'avaient pas été interrogés lors de l'enquête initiale, de sorte que la majorité des ménages interrogés pour l'enquête finale ont été ajoutés à ce stade. Une stratégie similaire a été adoptée au Maroc. En raison de la faible adhésion parmi les ménages identifiés comme emprunteurs potentiels, de nouveaux ménages ont été sélectionnés dans l'enquête finale, représentant 26 % de l'échantillon final. Si l'on tient également compte des taux d'attrition (disponibles uniquement pour les protocoles des panels) compris entre 8 % (Maroc) et 37 % (Mexique), il est clair qu'aucune des RCT n'a été menée conformément aux normes (échantillonnage non aléatoire des ménages ciblés en Bosnie-Herzégovine et en Mongolie, absence ou échec des panels pour les quatre autres RCT en raison de problèmes de collecte de données ou d'une faible adhésion).

Quoi qu'il en soit, il aurait été fondamental de vérifier l'équilibre des échantillons sur l'enquête initiale. Les études varient considérablement en termes de variables testées. Certaines ont testé un nombre étonnamment réduit de variables par rapport au large éventail de données collectées (Mexique). D'autres en ont testé bien davantage, mais toutes diffèrent quant aux variables testées. Dans certains cas, la plupart des variables incluent au moins quelques-uns des résultats pour lesquels l'impact a été mesuré dans l'enquête finale. Dans le cas du Maroc, toutefois, les tests d'équilibre n'ont été appliqués qu'à des sous-ensembles spécifiques des variables de résultats évaluées dans l'enquête finale (par exemple, les ventes pour les ménages de cultivateurs ou d'éleveurs de bétail, en lieu et place des ventes globales déclarées dans l'enquête finale). Dans notre réplication, nous avons constaté des déséquilibres importants et significatifs dans ces résultats. Les ménages du groupe de traitement ont tiré 22 % moins de ventes et de bénéfices de leur travail indépendant que les ménages du groupe de contrôle (significatif au niveau de 5 %). Ils ont également investi 61 % de plus (significatif au niveau de 5 %). En outre, il existe dans l'enquête initiale des déséquilibres concernant un certain nombre de variables importantes, telles que la superficie des terres possédées, l'accès aux services de base et l'autonomisation des femmes. Outre les variables testées, la base de calcul a aussi son importance. Par exemple, l'étude mexicaine a limité ses tests d'équilibre à 1 823 ménages interrogés, tant dans l'enquête initiale que dans l'enquête finale. Or, si l'on pratique les mêmes tests sur tous les ménages enquêtés en *baseline* (6 786), comme cela a été fait en Inde et au Maroc, on constate des différences significatives dans le revenu des ménages par adulte au cours du mois précédent, en particulier chez ceux faisant partie d'un groupe informel[12].

Même si les différences de base entre les groupes de traitement et de contrôle ne sont pas statistiquement significatives, elles peuvent être considérables. En Mongolie et en Éthiopie, les tests d'équilibre en enquête initiale ont révélé des différences moyennes souvent supérieures à 10 % (jusqu'à 50 %), mais non significatives (ce qui n'est pas surprenant étant donné la petite taille des

12. Les calculs sont disponibles sur demande auprès des auteurs.

échantillons). Ils sont systématiquement interprétés au nom de ce qui semble être un souci de commodité (absence de déséquilibres et donc succès du processus de randomisation), alors que l'explication inverse est souvent avancée pour les résultats : lorsque les coefficients ne sont pas significatifs en raison du manque de puissance statistique, ils sont interprétés comme étant « économiquement significatifs ».

Aucun des articles n'aborde en profondeur les questions d'erreur de mesure. Pourtant, la littérature insiste sur la difficulté d'obtenir des mesures fiables pour bon nombre des résultats analysés, en particulier la consommation des ménages et la production des micro-entreprises et de l'agriculture (Deaton, 1997 ; Grosh et Glewwe, 2000). Les erreurs de mesure sont simplement mentionnées dans une note de bas de page consacrée au biais de mémoire potentiel dans le cas de l'Inde et dans une discussion sur la sous-déclaration des emprunts dans le cas du Maroc, supposée expliquer les différences entre les données administratives et les enquêtes. Seule la RCT éthiopienne émet de sérieux doutes sur la qualité des données et reconnaît explicitement que ce problème affecte la validité interne. La RCT mexicaine précise que l'enquête initiale a dû être interrompue et que ses données n'ont pas pu être utilisées parce qu'elles n'étaient pas fiables, sans donner de détails ni expliquer comment des données plus fiables auraient pu être recueillies dans l'enquête finale. Malheureusement, il apparaît impossible de traiter plus en détail la qualité des données à partir des seuls articles. Cependant, une analyse détaillée de la cohérence des données et du recodage pratiqué par des chercheurs dans le cas marocain (Bédécarrats *et al.*, 2019a) montre que ce problème a altéré les résultats. Certains éléments tendent à indiquer que des problèmes similaires peuvent tout à fait se présenter dans d'autres cas. Par exemple, une analyse préliminaire des données mexicaines révèle que les tranches d'âge ne correspondent pas entre les enquêtes pour 231 (12,7 %) des 1 823 femmes interrogées, à la fois dans l'enquête initiale et dans l'enquête finale de Banerjee *et al.* (2015b).

Validité externe

La question de la validité externe des RCT est la plus débattue dans la littérature sur ce sujet. La validité externe est une question primordiale, d'autant plus que, contrairement à beaucoup de données observationnelles, les RCT sont menées à petite échelle et dans des lieux non représentatifs, comme on l'a vu plus haut. La validité externe est également menacée lorsque l'échantillonnage est sélectif, c'est-à-dire lorsqu'une étude se concentre sur des sites et des catégories de population spécifiques. Intervient ensuite le biais lié aux opérateurs, par exemple, lorsque les résultats obtenus par une ONG diffèrent d'une même intervention réalisée à plus grande échelle par un gouvernement (Bold *et al.*, 2013 ; Vivalt, 2020). Or, la question de la validité externe des RCT est rarement prise en compte par les *randomistas*. Peters *et al.* (2018) ont systématiquement passé en revue toutes les (54) RCT publiées dans les principales revues économiques de 2009 à 2014 afin d'évaluer les menaces les plus préoccupantes pesant sur la validité externe (effets

Hawthorn/Henry[13], effets d'équilibre général, problèmes d'échantillons spécifiques et attention accordée à l'offre de traitement). Sur la base d'un ensemble d'indicateurs objectifs, mais de critères statistiquement moins exigeants, l'article constate que la majorité des RCT publiées passent ces menaces sous silence et que beaucoup ne donnent pas les informations nécessaires pour évaluer les problèmes potentiels.

La validité externe a aussi un lien avec la pertinence des résultats sélectionnés. Se concentrer sur l'impact « moyen » et la difficulté à traduire l'hétérogénéité des impacts et leur distribution constituent un obstacle majeur à la pertinence des résultats (Ravallion, 2009a ; DFID, 2012 ; Vivalt, 2020). Le fait de se cantonner à un impact à court terme (pour des raisons de coût et d'attrition) signifie souvent que l'on étudie des indicateurs à mi-parcours, qui peuvent se révéler très différents des résultats finaux (Boone *et al.*, 2013), si ce n'est totalement inverses, puisque les trajectoires de projets sont rarement linéaires (Labrousse, 2010 ; Woolcock, 2013). Pourtant nombreux, les effets d'entraînement et d'équilibre général sont occultés, même si ce n'est que partiellement le cas pour la RCT marocaine (Ravallion, 2009a ; Acemoglu, 2010 ; Deaton et Cartwright, 2018). Il en va de même pour les considérations politiques afférentes à la réplication des programmes, en dépit de leur importance pour la transposition à plus grande échelle (Acemoglu, 2010 ; Bold *et al.*, 2013 ; Pritchett et Sandefur, 2013b). Enfin, point essentiel, les *raisons* de l'impact sont ignorées : les RCT peuvent fort bien permettre de mesurer et de tester certains impacts et aspects des interventions, mais elles ne peuvent analyser ni leurs *mécanismes* ni leurs *processus* sous-jacents. Nonobstant les limites de la méthode, l'absence de théorie empêche toute forme de compréhension des processus de changement. Pour pallier cette limitation de la théorie probabiliste de la causalité, il faudrait un « modèle causal » (Cartwright, 2010), une théorie cohérente du changement (Woolcock, 2013), une approche structurelle (Acemoglu, 2010) et une évaluation de l'intervention dans son contexte (Ravallion, 2009a ; Pritchett et Sandefur, 2015).

Le tabl. 3 résume les problèmes de validité interne tels qu'ils peuvent être évalués à l'appui des informations dont nous disposons. Les lacunes usuelles des RCT demeurent valables ici.

En premier lieu, l'échantillonnage est sélectif : les critères de sélection de l'expérience sont *ad hoc*, puisque les RCT ont été conduites dans des zones d'expansion des IMF. Comme le démontre Wydick (2016), la contrainte de la randomisation (identification de populations ou de zones vierges) a obligé les *randomistas* à choisir des populations et des zones « marginales » jusque-là négligées par les IMF et donc très spécifiques par rapport au marché « normal ». Les tentatives infructueuses menées par les études marocaines, mexicaines et indiennes pour

13. Il s'agit de biais comportementaux induits par l'expérience dès lors que les sujets savent qu'ils y participent : biais sur le groupe de traitement (effet Hawthorne) ou sur le groupe de contrôle (effet John Henry). Dans le domaine médical, les RCT en simple ou double aveugle (sujets et expérimentateurs) sont ordinairement utilisées pour contrôler ces biais (voir Abramowicz et Szafarz, chap. 10, ce volume).

identifier des emprunteurs potentiels montrent qu'il est difficile de caractériser la population cible du microcrédit. Cela exclut en toute légitimité la possibilité d'extrapolation à une population plus large. *A fortiori*, les échantillons retenus ne sont représentatifs de rien, si ce n'est d'eux-mêmes : les ménages interrogés dans le cas de la Bosnie-Herzégovine et de la Mongolie, et les zones d'expansion (villages et quartiers sélectionnés) dans le cas des quatre autres pays. De plus, cette propriété n'existe qu'en théorie : les multiples défaillances des protocoles d'enquête sur le terrain font que les échantillons théoriquement représentatifs des zones d'expansion ne le sont pas *de facto*.

Si les données ne peuvent pas être extrapolées, la comparaison avec d'autres sources peut se révéler utile pour qualifier les profils des personnes interrogées. Les chiffres officiels provenant d'enquêtes représentatives conduites par des offices nationaux des statistiques constituent un bon point de repère pour caractériser un contexte national ou local. Seules deux études l'ont fait (Bosnie-Herzégovine et Mongolie). Dans les quatre autres études, il est difficile de savoir qui sont les personnes interrogées. Comme indiqué ci-dessus, nous avons effectué cet exercice pour la RCT marocaine. Nous avons démontré, entre autres résultats, que la taille moyenne des ménages est atypique et tend à augmenter, alors qu'elle décroît dans le reste de la population sur la même période. Pour aller plus loin dans cette évaluation, nous utilisons la typologie des risques sur la validité externe établie par Peters *et al.*, (2018) : effets Hawthorne et John Henry et effets d'équilibre général (les autres étant abordés ci-dessus). Les articles ne traitent pas de ces risques et beaucoup ne fournissent pas les informations nécessaires pour évaluer les problèmes potentiels, à l'exception (partielle) des effets Hawthorne (Bosnie-Herzégovine et indirectement Mexique, voir la discussion sur l'éthique ci-dessous) et des effets d'équilibre général et d'entraînement (Maroc), bien que ces effets soient à l'œuvre dans tous les cas.

Ces risques pour la validité externe et interne sont-ils reconnus par les *randomistas* ? Plus généralement, quels types de réserves formulent-ils dans leurs articles ? Nous en rendons compte dans le tabl. 3. À l'exception de la RCT marocaine, les auteurs formulent un certain nombre de réserves. Presque tous mentionnent le manque de validité externe induit par le manque de puissance statistique, lui-même dû à la taille insuffisante des échantillons. De même, l'hétérogénéité des traitements est largement reconnue. Le fait que les autres RCT donnent des résultats similaires (mais tout aussi insuffisants en termes de puissance statistique) est perçu comme une source de robustesse (voir, par exemple, Banerjee *et al.*, 2015b : 25). En outre, des réserves plus spécifiques sont citées, notamment au regard du non-respect du plan de sondage (Éthiopie), de l'attrition sélective (Inde et Éthiopie) et des erreurs de mesure (Éthiopie). Ces observations tendent à confirmer la persistance des résultats de Peters (2018) concernant l'attention limitée accordée à la validité externe, ce à quoi il faut ajouter les problèmes de validité interne évoqués ci-dessus.

Enfin, les considérations éthiques méritent d'être discutées, car elles revêtent une dimension spécifique pour les RCT en général (voir l'introduction, ce volume ; Ravallion, chap. 1, ce volume ; et Abramowicz et Szafarz, chap. 10,

ce volume). Or, contrairement à toute attente, aucun des articles ne fait état de ces considérations. Ils ne précisent pas si le consentement éclairé des participants a été demandé et obtenu, à l'exception d'ANGELUCCI *et al.* (2015) sur Compartamos au Mexique. En outre, les informations qu'ils déclarent avoir communiquées aux participants sont partielles : ils précisent, peut-être pour écarter tout soupçon d'effet Hawthorne, qu'ils leur ont demandé leur accord pour participer à une « enquête de recherche socio-économique approfondie ». Mais ils ont sciemment omis de mentionner que l'enquête était en lien avec Compartamos et, surtout, qu'elle faisait partie d'une expérimentation. L'examen des questionnaires d'enquête disponibles (Bosnie-Herzégovine et Maroc) montre que, dans ces deux cas, les répondants n'ont pas été informés qu'ils participaient à une expérimentation. La RCT bosniaque soulève d'autres questions éthiques. Cette RCT consistait à accorder des crédits à des personnes qui, selon les critères de solvabilité de l'IMF, avaient initialement été refusées, comme en Afrique du Sud et aux Philippines (KARLAN et ZINMAN, 2009 ; 2011). Cette stratégie faisait peser un risque sur le groupe traité, au mépris du principe de « ne pas nuire ». La RCT confirme que les clients marginaux ont beaucoup plus de mal à rembourser que les clients réguliers, avec un risque de surendettement[14].

Maintenant que nous avons discuté des questions de validité interne et externe, penchons-nous sur la question des impacts eux-mêmes. Même si l'on ne prend pas en compte les limites exposées ci-dessus et que l'on s'en tient aux résultats proposés par les auteurs, les impacts sont problématiques. Le tabl. 5 en donne un aperçu. Premièrement, les données sur l'adhésion ne sont pas fiables et présentent souvent des contradictions entre les sources d'enquête et les sources administratives. Le cas marocain montre que les incohérences vont au-delà des différences de moyennes et de la sous-déclaration (voir la section « Le cas emblématique de la RCT marocaine » sur les écarts entre les données administratives et les données d'enquête). En moyenne, l'impact des expériences sur l'adhésion du crédit varie de 8 à 50 %, lorsque les groupes ont été randomisés, à 98,5 % en Bosnie-Herzégovine, où ce sont les individus qui ont été randomisés.

En ce qui concerne les impacts sur le microcrédit, le faible taux d'adhésion a de considérables implications sur le niveau de significativité des coefficients estimés. DAHAL et FIALA (2020) répliquent les six RCT de l'*AEJ:AE*. Ils constatent que chacune d'entre elles souffre d'un important manque de puissance statistique en raison de la faible adhésion du produit financier offert. Même après la mise en commun des données, les magnitudes minimales des effets détectables demeurent très élevées : 230 % pour les principaux résultats en conformité parfaite et 1 000 % en conformité effective. Ils concluent dans leur résumé que « les études existantes sur l'impact de la microfinance pâtissent généralement d'un manque de puissance statistique pour identifier les impacts de manière fiable et suggèrent que nous en savons encore très peu sur l'impact de la microfinance ». Quoique BANERJEE *et al.* (2015b) admettent le problème de la sous-puissance statistique

14. « Tout cela indique que les agents de crédit avaient de bonnes raisons de classer comme marginale notre population cible » (AUGSBURG *et al.*, 2015 : 201).

dans leur introduction, l'article de DAHAL et FIALA (2020) est le premier à quantifier l'ampleur du problème. Il confirme l'étude précédente de MCKENZIE (2012), qui estime à 15 000 000 la taille d'échantillon nécessaire pour pouvoir garantir la capacité à déceler des magnitudes d'impact de 10 % dans la RCT indienne.

Pour ce qui est des impacts sur les résultats sélectionnés, la présentation faite à ce sujet par les auteurs de l'introduction générale (tabl. 2), qui est supposée résumer les résultats consolidés des six RCT, est trompeuse. Un décompte exhaustif des impacts estimés sur toutes les variables examinées dans les six articles permet de tirer les conclusions suivantes. Pas moins de 298 impacts sont estimés sur l'ensemble du volume (sans compter les estimations des quantiles). Sur ce total, seuls 10 sont significatifs au niveau de 1 %, ce qui signifie que 97 % des effets possibles retenus ne sont pas significativement différents de zéro. Trois RCT n'ont aucun impact significatif (Bosnie-Herzégovine : 0/47, Éthiopie : 0/37 et Mongolie : 0/41) et l'une d'entre elles n'a qu'un seul impact significatif (Inde : 1/99). Même lorsque le seuil est abaissé à 10 % (un niveau moins rigoureux que dans la pratique habituelle), 81 % des effets ne sont pas significatifs. La RCT bosnienne est un cas extrême à cet égard, avec seulement trois impacts significatifs à ce seuil sur les 47 testés. Ces proportions suscitent d'autant plus de doutes que tous les articles dénoncent un problème systématique de sous-puissance statistique, ce qui expliquerait l'absence d'impact. La taille des échantillons n'est pas suffisante pour estimer les impacts, compte tenu du faible taux d'adhésion, et c'est effectivement ce que nous constatons. De plus, 60 % des impacts significatifs (à 1 %) proviennent de la RCT marocaine, alors qu'elle représente à peine 12 % du nombre total d'impacts estimés. Ce résultat confirme le rôle central joué par cette expérience dans le numéro spécial, indépendamment de ses vertus supposées concernant sa stratégie novatrice d'échantillonnage et sa tentative pionnière d'estimation des effets d'entraînement. Cependant, nous avons montré le caractère douteux des résultats obtenus par cette RCT. Cela réduit encore le nombre d'impacts significatifs, qui était déjà notablement faible.

Symptomatiquement, le cheminement des résultats des articles académiques, d'abord dans l'introduction générale, puis dans la synthèse proposée dans le *Policy Bulletin* (J-PAL et IPA, 2015) participe, par des approximations successives, de la simplification et de la généralisation abusive des leçons retenues, voire de la présentation de résultats erronés. Si nous revenons au résumé des impacts présenté dans le *Policy Bulletin* (p. 11, tabl. 2 ; voir aussi notre tabl. 2), sur les 48 impacts mesurés (8 résultats et 6 pays), 16 sont annoncés comme significatifs (14 positifs et 2 négatifs). On est loin du compte. Premièrement, le seuil de significativité choisi est de 10 %, ce qui correspond à un niveau de précision situé à la limite supérieure de celui qui est habituellement utilisé. Si nous adoptons un seuil plus exigeant et plus proche des pratiques habituelles (à savoir 1 %), aucun des 16 impacts n'est significatif.

Une analyse plus détaillée des 16 impacts sélectionnés révèle de nombreuses incohérences. Pour la Bosnie-Herzégovine, les impacts sur la *propriété des entreprises* et sur les *stocks/actifs des entreprises* sont annoncés comme positifs. Mais le premier n'est pas significatif à 10 %. Quant au second, ce qui est

significatif à 10 % est une variable fictive qui mesure si l'entreprise détient ou non du capital. L'impact sur la valeur totale des *actifs* est négatif (quoique non significatif), donc au mieux nul. Pour l'Éthiopie, le seul impact considéré comme significatif et négatif est celui sur les *dépenses/consommation des ménages*. Or, la consommation n'a pas été mesurée dans l'enquête. En Inde, les deux impacts positifs sont sur les *stocks/actifs des entreprises* et les *stocks/coûts des entreprises*. Aucun de ces deux impacts n'est robuste : le premier impact est positif dans la deuxième enquête finale, mais n'est pas significatif dans l'enquête initiale, et inversement pour l'impact sur les *stocks/coûts des entreprises*. Au Mexique, on note deux impacts positifs. Si l'impact se vérifie pour les *revenus des entreprises*, aucune donnée ne permet de mesurer l'*investissement* (ce deuxième résultat étant supposé augmenter avec le traitement). Les *actifs* sont en outre en baisse (effet significatif à 5 %). En Mongolie, trois résultats sont censés avoir des effets positifs. Cette conclusion vaut pour deux d'entre eux : *propriété des entreprises* et *consommation des ménages* (à 10 %). Cependant, bien que l'indice composite des *actifs* soit positivement impacté (à 10 %), l'effet n'est pas significatif (et même négatif) pour la valeur des *actifs*. Dans le cas du Maroc, où quatre résultats sont jugés positifs, nous renvoyons aux problèmes de fiabilité de cette RCT mentionnés ci-dessus. La synthèse du *Policy Bulletin* semble biaisée, ou au mieux très imprécise.

Compte tenu de ces lacunes, les coefficients élevés, mais non significatifs auraient été identiques même si la taille des échantillons s'était révélée suffisante. Ces résultats ont deux implications. Premièrement, ils remettent en question le postulat général selon lequel le microcrédit n'est pas « transformateur ». C'est peut-être vrai, mais finalement aucune donnée fiable ne permet de le prouver. Deuxièmement, Dahal et Fiala (2020) concluent que « les études existantes […] suggèrent que nous en savons encore très peu sur l'impact de la microfinance »[15]. Ce paradoxe, compte tenu des ressources consacrées aux RCT sur le microcrédit, est confirmé par Morduch (2020), l'un des meilleurs spécialistes au monde du microcrédit.

Une autre conclusion concernant la validité externe et interne concerne le fait qu'aucune des études de réplication (Kingi *et al.*, 2018 ; Meager, 2019 ; Dahal et Fiala, 2020) n'ait pointé les erreurs que nous avons documentées dans notre réplication marocaine, notamment les plus évidentes, comme les déclarations des auteurs sur l'absence totale de contamination dans les groupes de contrôle, les comptages incohérents des ménages avant et après le *trimming*, et l'affirmation selon laquelle aucun *trimming* n'a été effectué dans l'enquête initiale. Cela ne fait que souligner les lacunes des réplications « *push-button* » ou des réplications qui appliquent des spécifications économétriques différentes aux mêmes données sans vérifier la fiabilité des données, des codages ou de l'échantillonnage d'origine.

15. Ce point est reconnu de façon détournée par les éditeurs du numéro spécial : « Les études individuelles peuvent manquer de preuves solides pour démontrer les effets transformateurs sur l'emprunteur moyen, mais elles manquent également de preuves solides contre les effets transformateurs » (Banerjee *et al.*, 2015c : 3).

Tableau 5
Impact, références et publications.

	Bosnie-Herzégovine	Éthiopie	Inde	Mexique	Mongolie	Maroc
Impacts						
Adhésion au crédit IMF						
Source des données	Enquête	Enquête	Enquête	Admin., enquête	Enquête	Admin., enquête
Présence d'autres IMF	Oui	Oui	Oui	Oui	Oui	Oui
Effet de substitution/diminution d'autres formes de crédit	n. a.	Non	Oui (substitution)	Oui (diminution)	Oui (substitution)	Oui (substitution)
Impact	Positif (98,5 %)	Positif (25 %)	Positif (13 %)	Positif 8 % (enquête), 11 % (admin.)	Positif (50 %)	Positif 9 % (enquête), 17 % (admin.)
Résultats (autres que l'adhésion au crédit)						
Nombre	47	37	99	37	41	37
Nombre d'impacts significatifs (à 1 % ; 10 %)	0/47 (1 %) 3/47 (10 %)	0/37 (1 %) 5/37 (10 %)	1/99 (1 %) 13/99 (10 %)	3/37 (1 %) 9/37 (10 %)	0/41 (1 %) 10/41 (10 %)	6/37 (1 %) 17/37 (10 %)
Références, publications						
Nombre de références :	22	24	27	28	37	16
Dont RCT	5	11	11	20	10	8
Dont méthodologie/théorie	4	4	3	4	18	6
Dont autres méthodes de microcrédit	6	7	4	4	5	0
Autres	7	2	9	0	4	2
Nombre d'articles dans des revues académiques	1 *(AEJ:AE)*	2 *(AEJ:AE ; Demography)*	1 *(AEJ:A)*	1 *(AEJ:AE)*	1 *(AEJ:AE)*	1 *(AEJ:AE)*

Source : Florent Bédécarrats, Isabelle Guérin et François Roubaud, sur la base de Banerjee *et al.* (2015c).

Notes : le nombre élevé de résultats dans la RCT indienne (99) est dû au fait que deux enquêtes finales ont été réalisées. Impact significatif (à 1 % ; 10 %) : nombre d'estimations d'impact associées à des valeurs *p* significatives au niveau de 1 % et au niveau de 10 %. IMF : institution de microfinance.

Résultats : des biais statistiques aux biais interprétatifs

La partie « Validité et portée du numéro spécial : évaluation critique » a exploré la fabrique des RCT sur le terrain et mis en exergue les nombreuses lacunes générées par des problèmes de validité interne et externe. Aux étapes de la collecte des données statistiques et de l'analyse économétrique fait suite l'étape de l'interprétation : « La beauté des évaluations randomisées est que les résultats sont ce qu'ils sont : nous comparons le résultat dans le traitement avec celui dans le groupe de contrôle, nous voyons s'ils sont différents et, si oui, de combien » (Banerjee, 2007 : 115-16). Une analyse de la manière dont les *randomistas* transforment leurs données en déclarations scientifiques remet en question cette prétendue « beauté » des RCT.

Pris isolément, la plupart des résultats économétriques des six RCT n'ont aucun sens en eux-mêmes, sans parler de l'absence d'informations contextuelles. Les auteurs, notamment dans l'introduction générale, font cette interprétation dans un contexte très spécifique et au prix d'hypothèses implicites, mais puissantes, empruntées à une théorie comportementale du changement. Un cadre d'anthropologie et un cadre d'économie politique donneraient lieu à des conclusions très différentes. Notre but n'est pas de discréditer le processus d'interprétation, qui est inhérent à l'analyse des données, mais de démontrer que les *randomistas*, contrairement à ce qu'ils prétendent, ne peuvent y échapper. Les résultats ne sont pas « ce qu'ils sont », comme le montre également Kabeer (2019) en employant des outils qualitatifs pour revisiter un terrain étudié par une RCT.

De plus, leur interprétation est fondée sur une « rhétorique de persuasion » (Labrousse, chap. 8, ce volume), qui consiste à faire table rase des recherches antérieures et à extrapoler (ici se pose à nouveau le problème de la validité externe), tout en faisant abstraction des questions spécifiques qui sont essentielles pour comprendre les impacts du microcrédit et que d'autres méthodes ont déjà abordées.

Faire table rase des recherches antérieures

Les résultats des *randomistas* sont souvent présentés comme des « découvertes » sans précédent, alors qu'ils ne sont souvent que la réplique de conclusions tirées d'études antérieures, notamment celles formulées à partir de méthodes non expérimentales qui ne sont presque jamais citées (Labrousse, 2010). L'introduction générale en est une bonne illustration. D'après la présentation qui en est faite, les résultats constituent les premières preuves scientifiques des impacts du microcrédit. « La base de preuves pour consacrer le microcrédit était assez mince » (Banerjee *et al.*, 2015c : 1). Jusqu'à présent, les preuves empiriques disponibles étaient basées sur « des anecdotes, des statistiques descriptives ou des études d'impact qui ne permettent pas de distinguer la causalité de la corrélation » (*ibid.* : 1-2). Les auteurs prétendent prendre part aux « débats qui ont eu lieu dans les années 2000 et qui perdurent aujourd'hui » (*ibid.* : 2), mais,

de fait, ces débats se déroulent dans un cercle étonnamment cloisonné. Sur les 18 références mentionnées dans l'introduction générale, 12 (les deux tiers) proviennent des auteurs eux-mêmes et 17 (94,4 %) de membres du J-PAL. Un seul article échappe à ce principe endogamique.

Aucune étude non randomisée n'est citée. Pour ce qui est des six articles du numéro spécial, l'article sur le Maroc est tout aussi exclusif (seules des RCT sont mentionnées). Les autres le sont moins, mais de manière variable, comme le montre le tabl. 5. L'étude sur la Bosnie-Herzégovine est la plus pluraliste, avec un rapport RCT/non RCT de 0,8 ; ce rapport varie de 1,57 à 5 pour les autres.

Ce mépris des preuves disponibles non issues de RTC s'accompagne d'une tendance à extrapoler et à négliger les questions clés. Sans prétendre à l'exhaustivité, mais en se concentrant sur les points qui nous paraissent essentiels, nous abordons tour à tour les questions d'adhésion, de la création d'entreprises et de la liberté de choix, des transferts sociaux et de l'autosuffisance, ainsi que le problème du surendettement.

Adhésion

La faible adhésion constitue certainement le résultat le plus abouti du numéro spécial. De nombreux praticiens, décideurs et chercheurs s'échinent aujourd'hui encore à prédire un marché illimité, confondant ainsi exclusion financière et demande de crédit. Bien que ce résultat soit utile, sa véritable signification demeure limitée. Tout d'abord, il convient de noter que cet exercice n'a rien de nouveau. Certaines études mettent depuis longtemps en garde contre la faiblesse de la demande de microcrédit (JOHNSON et ROGALY, 1997 ; SERVET, 2006), notamment en fournissant des estimations quantitatives (KHANDKER *et al.*, 1998 ; HES et POLEDŇÁKOVÁ, 2013). En outre, les taux d'adhésion évoqués ici sont difficiles à comparer et à interpréter, étant donné la diversité des protocoles et des méthodes de randomisation (voir la section « Validité externe »). Il apparaît donc difficile d'évaluer la nature et la signification de la population cible et, par conséquent, de tirer des conclusions opérationnelles. De surcroît, les RCT ne disent rien sur les raisons de la faible adhésion : est-elle le reflet d'une demande intrinsèquement modeste et d'une faible propension à l'endettement et/ou d'une inadéquation de l'offre, étant entendu que les deux explications ne sont pas mutuellement exclusives ? Seules des données plus détaillées pourraient répondre à cette question, supposant une analyse fine des pratiques financières, comme celle des « agendas financiers » (*financial diaries*) (COLLINS *et al.*, 2009) et leurs implications sociales, morales et politiques (voir, par exemple, l'analyse qualitative du contexte marocain, négligé par les auteurs de la RCT marocaine : MORVANT-ROUX *et al.*, 2014).

Microcrédit, travail indépendant et liberté de choix

Les six études du numéro spécial tendent à s'accorder sur le fait que les impacts sur la création d'entreprises sont limités (significatifs dans deux cas seulement), l'expansion des entreprises existantes étant un impact plus fréquent (quatre cas).

Une amélioration de la rentabilité n'est constatée que dans un seul cas (Maroc), mais nous avons constaté plus haut que la validité interne de ces résultats laisse à désirer. En outre, même dans le cas d'une création ou d'une expansion d'entreprise, on n'observe aucun impact sur la croissance des revenus, soit parce que la rentabilité est faible, soit parce que les revenus du travail indépendant sont compensés par une baisse de revenus d'emplois salariés. Les auteurs de l'introduction générale prétendent ainsi tirer une conclusion inédite sur l'impact du microcrédit sur l'entrepreneuriat.

Cependant, depuis la fin des années 1980, de nombreuses études empiriques ont été menées pour mesurer l'impact du microcrédit[16]. L'étude systématique de Duvendack *et al.* (2011), conduite alors que les RCT commençaient à peine à voir le jour, tire deux conclusions. Premièrement, beaucoup d'études quantitatives, tant expérimentales (RCT comprises) qu'observationnelles, sont sujettes à de multiples biais[17]. Deuxièmement, lorsque les résultats sont valables, ils révèlent un impact limité et hétérogène, ce que Morduch a également observé à la fin des années 1990 dans son article pionnier sur les promesses partiellement non tenues du microcrédit (Morduch, 1999). Les résultats du numéro spécial n'ont donc rien d'inédit. Plus important encore, étant donné la complexité des chaînes causales induites par le microcrédit (Duvendack *et al.*, 2011) et l'hétérogénéité des effets et des types de microcrédit[18], les RCT ne paraissent pas adaptées (Bernard *et al.*, 2012). Enfin, la question des *randomistas* – le microcrédit fonctionne-t-il ou non ? – est mal posée. Ce que montrent les études rigoureuses (qu'elles soient quantitatives, qualitatives ou mixtes), c'est que certains types de microcrédit peuvent être utiles à certaines catégories de populations et dans certains contextes, mais pas d'autres (Bédécarrats, 2012 ; Copestake *et al.*, 2016). Par exemple, les travaux de Copestake *et al.* (2001 ; 2005) au Pérou et en Zambie et de Bouquet *et al.* (2007) à Madagascar indiquent précisément quelles catégories de population bénéficient du microcrédit, et pourquoi, et, inversement, quelles catégories voient leur situation se dégrader, avec des propositions opérationnelles directes sur l'amélioration des services proposés afin d'en augmenter l'impact. Toujours à Madagascar et dix ans avant le numéro

16. Bédécarrats (2012) a identifié 154 études d'impact, contre 51 pour Duvendack *et al.* (2011).

17. Plusieurs réplications d'études non expérimentales longtemps prises comme « preuves » de l'impact positif du microcrédit ont révélé de nombreux biais et une surestimation des impacts. Voir Duvendack et Palmer-Jones (2012) ; Roodman et Morduch (2014).

18. Nous donnerons l'exemple du microcrédit rural, qui est largement représenté dans le numéro spécial. Au-delà des modalités du crédit, quels sont les besoins en crédit (intrants, équipements, bétail, trésorerie pour financer la saison creuse, etc.), de quel type d'agriculture parlons-nous (cultures commerciales ou vivrières, agriculture en zone sèche ou pluviale, intensive ou extensive, familiale ou professionnelle, indépendante ou contractuelle par intégration dans des secteurs agro-industriels ou des coopératives de producteurs, etc.) et quelle est la nature des économies rurales (degré de monétarisation, éloignement et qualité des infrastructures, opportunités de revenus non agricoles) ? Et, surtout, de quel type d'IMF parlons-nous ? Le statut (but lucratif/non lucratif) est certes un facteur de distinction (spécifié dans le numéro spécial), mais d'autres questions essentielles se posent, notamment le mode de gouvernance, le degré d'intégration et d'adaptation aux réalités locales et la capacité à concevoir des produits adaptés à la demande locale. Compte tenu de cette diversité, il est absurde de parler de « microcrédit rural ». Sur cette diversité, voir par exemple Morvant-Roux (2009).

spécial, notre propre évaluation de l'impact d'une IMF locale basée sur une approche quasi expérimentale suggère trois grandes caractéristiques stylisées, présentées plus tard comme des « découvertes » par les *randomistas*. L'impact du microcrédit n'est pas « transformateur », les impacts sont hétérogènes au regard de la distribution des entreprises par taille et le contexte importe : le microcrédit est plus bénéfique en période de croissance qu'en période de crise (Gubert et Roubaud, 2011).

Pour comprendre l'hétérogénéité des impacts (et en tirer des conclusions opérationnelles), une conception distincte des mécanismes de causalité est nécessaire : le but n'est pas de « faire la différence », mais de penser en termes de « mécanisme » et de « processus » (Shaffer, 2015). De plus, étant donné les nombreuses externalités, il apparaît tout aussi restrictif de se concentrer sur l'impact individuel. Très peu d'études ont appliqué des modèles d'équilibre général au cas du microcrédit à méso-échelle (pour une exception, voir Mahjabeen, 2008). Les examens des externalités ont été principalement réalisés par des analyses d'économie politique, puisque c'est précisément l'analyse de l'ancrage des IMF dans leur environnement social, culturel, politique et économique et des externalités qui a un puissant effet sur l'adhésion des produits et, partant, sur leur impact (Copestake *et al.*, 2016). Des études d'impact convaincantes et utiles conduites dans des zones rurales ont démontré le rôle clé des innovations financières ancrées dans les territoires locaux – capables de développer des produits spécifiques conçus à l'échelle locale (crédits-relais, fonds de garantie et crédit-bail) et associés à d'autres mesures (contrats de métayage, entrepôt à récoltes, assistance technique, etc.) – pour inciter les petits agriculteurs à participer à différentes chaînes de valeur (Bouquet *et al.*, 2007 ; Bastiaensen et Marchetti, 2011), tout en rencontrant souvent des effets de seuil (Doligez, 2002). Les effets sont parfois discutables, comme lorsque le microcrédit accélère les processus de migration, puisque la migration devient nécessaire pour rembourser les microcrédits (Morvant-Roux, 2013 ; Bylander, 2014). Leur nature est parfois plus politique et culturelle qu'économique. En Égypte, par exemple, l'introduction du microcrédit perturbe les valeurs locales – définies au sens large comme ce qui fait sens pour les gens – et, par conséquent, les processus de reconnaissance, d'identité et de socialisation (Elyachar, 2006). Dans l'Inde rurale du Sud, la présence massive d'IMF dans certains territoires reconfigure les rapports de force locaux et les chaînes de patronage en les féminisant (Guérin et Kumar, 2017). Ces résultats (et les questions qui s'y rapportent) sont très éloignés de ceux des *randomistas*. Et pourtant, si nous voulons vraiment comprendre ce que le microcrédit change dans la vie des gens, c'est précisément ce type de questions larges qu'il faut nous poser.

Outre ces études approfondies, qui reposent systématiquement sur une bonne connaissance des contextes locaux dans le temps, il est utile de mentionner d'autres méthodes plus légères, conçues pour identifier rapidement les caractéristiques des clients (et des non-clients) et leur mode d'utilisation des services et en tirer des recommandations en vue d'améliorer la qualité de l'offre, qui

demeure la principale question récurrente posée par les fournisseurs de microcrédit[19].

Revenons maintenant au numéro spécial. Non seulement les auteurs n'ajoutent rien de fondamentalement nouveau aux débats, mais leur interprétation des résultats quantitatifs pose problème. La micro-entreprise peut refléter l'absence de choix, et non la multiplication des choix. Une grande partie des micro-entrepreneurs, condamnés au travail indépendant faute d'emploi rémunéré, relève plus de l'auto-exploitation analysée par CHAVANOV (1966 [1925]) que de l'entrepreneuriat schumpétérien (LAUTIER, 2004). Le cas de la Mongolie est instructif à cet égard. La RCT montre que l'accès au crédit solidaire permet aux femmes de créer de nouvelles micro-entreprises, mais pour des revenus négatifs, alors que leur temps de travail augmente de plus d'un tiers (sans variation du temps passé au foyer). Ces effets négatifs sont principalement observés chez les femmes moins instruites (ATTANASIO *et al.*, 2015 : 105, note 21). Les auteurs pensent que la rentabilité peut s'améliorer une fois le crédit remboursé (ATTANASIO *et al.*, 2015 : 115). On retrouve ici le problème de la temporalité qui, comme nous l'avons déjà souligné, limite fortement l'intérêt des méthodes randomisées (LABROUSSE, 2010 ; BÉDÉCARRATS *et al.*, 2019b). Ces femmes peuvent en effet avoir choisi de se lancer dans l'aventure entrepreneuriale, ce qui peut expliquer l'amélioration de la consommation (les résultats indiquent une consommation plus importante et plus saine). Mais que signifie ce « choix » et, surtout, quelles en sont les conséquences s'il entraîne un accroissement des responsabilités et éventuellement un désengagement des autres membres du ménage (et donc des inégalités intrafamiliales) ? Les données quantitatives ne permettent pas de tirer une conclusion à ce sujet et les auteurs de la RCT ne portent pas de jugement particulier. Une interprétation robuste exigerait d'autres types de données, quantitatives ou qualitatives. Les auteurs de l'introduction générale, pour leur part, se concentrent uniquement sur la dimension de « liberté de choix », sans mentionner les effets potentiellement négatifs de ces « choix » sur les femmes, en particulier les plus défavorisées.

Microcrédit, dépenses sociales, transferts sociaux et autosuffisance

Si les effets en termes d'activité et de revenus ne sont pas concluants, les auteurs de l'introduction générale observent ce qu'ils décrivent comme des effets positifs sur deux indicateurs : les « dépenses non essentielles », signe d'une discipline accrue et de meilleures compétences de gestion, et la réduction des « transferts sociaux », signe d'une plus grande autonomie. Les « dépenses non essentielles » incluent les « biens de tentation » et ont diminué dans quatre pays (elles n'ont pas été mesurées en Éthiopie et les résultats n'étaient pas significatifs en Mongolie) :

19. Les outils élaborés par AIMS (Assessing the Impact of Microenterprise Services) et Imp-act, qui ont été dénigrés pour leur manque de méthode quantitative sophistiquée, en sont des exemples. Ces outils peuvent difficilement « prouver » l'impact à grande échelle, mais ils se sont révélés très utiles pour « améliorer » et diversifier l'offre de services de microfinance.

alcool et cigarettes en Bosnie-Herzégovine, cigarettes, sucreries et sodas au Mexique, alcool, tabac, feuilles de bétel, jeux de hasard et aliments consommés hors du foyer en Inde. Ces dépenses incluent également les festivités, avec des baisses observées en Inde et au Maroc.

Les auteurs avancent plusieurs explications de cette réduction des « biens de tentation » : contraintes de remboursement et d'investissement, meilleure autodiscipline et plus grande participation des femmes à la prise de décision. Le déclin des dépenses de tentation constitue l'un des principaux résultats de la RCT indienne présentés dans le résumé. Les auteurs de l'étude prennent soin de préciser que ce sont les populations elles-mêmes qui décrivent ces biens comme des « biens de tentation », en ce sens qu'elles souhaiteraient réduire ce type de consommation (Banerjee *et al.*, 2015b : 24). Mais le fait que les gens expriment cette préférence (une observation aux origines vagues, qui semble plutôt relever de l'« anecdote », dont l'utilisation est soulignée par Labrousse, chap. 8, ce volume), pourrait bien indiquer qu'ils ont intégré les discours moralisateurs fréquemment assénés par les organisations de développement (IMF comprises), et ce, depuis la période coloniale[20].

Au-delà de la dimension moralisatrice des conclusions des *randomistas*[21], une analyse détaillée de la signification et du rôle de ces dépenses pourrait apporter un éclairage différent. Sur le sujet de l'alcool, personne ne conteste que la consommation excessive pose un problème de santé publique. Pourtant, si nous voulons vraiment comprendre ce type de consommation et concevoir des lignes d'action, il est essentiel de reconnaître la dimension sociale et politique de l'alcool. Comme beaucoup d'autres biens de tentation, et contrairement à ce que suggère l'économie comportementale, il ne s'agit pas d'un bien qui se définit uniquement par son « utilité immédiate » (Banerjee et Mullainathan, 2010). L'alcool peut jouer un rôle social en ce sens qu'il permet aux travailleurs d'endurer un travail très exigeant physiquement et d'accéder à des espaces de socialisation, et donc à des informations stratégiques (les bars constituent souvent des lieux privilégiés pour négocier des contrats de travail et des commandes ; Picherit, 2018). L'alcool peut jouer un rôle politique lorsqu'il donne aux travailleurs l'occasion de formuler vis-à-vis des employeurs et des patrons des revendications qui sont plus facilement acceptables sous l'effet de l'ébriété (Scott, 1977). Et, surtout, l'alcool est souvent proposé délibérément par les employeurs et les recruteurs de main-d'œuvre afin de fidéliser les travailleurs (Picherit, 2018). Il est donc fallacieux de dire que le sacrifice ou une plus grande maîtrise de soi suffiraient pour lutter contre ces « tentations ».

20. En Inde, par exemple, les colons britanniques et les missions chrétiennes du début du XIX^e^ siècle dénonçaient déjà l'imprévoyance et la prodigalité des pauvres (Cederlöf, 1997 ; Hardiman, 2000).

21. Les discours des *randomistas* rappellent la morale victorienne de la révolution industrielle européenne, légitimée par les arguments des économistes néoclassiques de l'époque. Face à l'extrême pauvreté du monde ouvrier pendant la révolution industrielle britannique, certains ont déploré le manque d'autosuffisance des pauvres, leur manque de prévoyance et le gaspillage de leurs revenus en alcool, et ont plaidé pour des cours d'éducation financière plutôt que pour des hausses de salaire (voir par exemple Jevons, 1883 : 196-200, 205).

De même, les frais de restauration (repas et thé) hors du domicile ne sont pas uniquement des opportunités manquées « d'épargne lucratives » (Banerjee et Duflo, 2011 : 170). Les restaurants de rue et les gargotes de thé sont des lieux éminemment stratégiques. Dans une économie informelle opaque, structurée par des relations interpersonnelles, ces espaces permettent aux commerçants de se tenir mutuellement informés de la situation du marché, de l'évolution des prix, des opportunités à saisir, des sources de financement possibles, des risques de contrôles fiscaux ou policiers, etc. Les petits entrepreneurs y cultivent des liens d'échange et de soutien mutuel, qui ont une fonction souvent décisive pour la survie de leur entreprise.

En ce qui concerne les dépenses consacrées aux rituels sociaux et religieux, l'anthropologie a depuis longtemps démontré que la « richesse sociale » constitue un facteur essentiel de réussite et de protection (Guyer, 1997) et qu'« investir » dans les relations sociales peut, dans certaines situations, se révéler beaucoup plus rationnel que d'essayer d'économiser de l'argent en se coupant de son entourage (Narotzky et Besnier, 2014). Au-delà des *randomistas*, la question des « taxes communautaires » et de leur rapport coût-bénéfice en termes de protection a fait l'objet de diverses études par des économistes du développement. Mais ces études tiennent rarement compte de la complexité des circuits financiers auxquels ces dépenses donnent lieu et de leur caractère durable. Une analyse conduite en Inde sur la corrélation entre les dépenses festives et les invitations à déjeuner montre, par exemple, que ces dépenses font office de filet de sécurité (Rao, 2001). De plus, ce que les économistes considèrent comme une dépense est parfois vu comme une créance ou une épargne, puisqu'il s'ensuivra une compensation. Toujours en Inde, la comptabilisation de l'ensemble des dettes et des créances générées au fil du temps par les dépenses cérémonielles, dont les familles ont bien conscience puisqu'elles les calculent en ces termes, montre que la richesse financière nette des familles est radicalement différente de celle que suggère une analyse en termes de « dépenses » (Guérin *et al.*, 2019). Cela contredit l'idée avancée par les *randomistas* à propos des biais de court terme dont souffriraient les pauvres (Banerjee et Duflo, 2011 : 183-204).

En ce qui concerne les transferts sociaux, sur les huit estimations retenues (qui portent sur des transferts en provenance de la famille ou l'État), cinq sont négatives. Cette observation conduit les éditeurs du numéro spécial à conclure que l'« autosuffisance » s'est améliorée, un facteur qui est jugé de manière positive[22]. Cette interprétation est à la fois risquée – il n'y a aucune raison de croire que la baisse des transferts depuis la famille et les amis soit jugée positive ou considérée comme une source de bien-être par les personnes concernées – et normative, à l'instar des interprétations précédentes. Là encore, l'anthropologie est d'une aide précieuse pour élucider le rôle décisif des interdépendances

22. Il convient toutefois de noter que cette interprétation est celle des auteurs de l'introduction, et non des auteurs des articles, qui ne commentent pas ce résultat, mais en soulignent bien l'ambiguïté. Pour ce qui est de la Mongolie, les auteurs mentionnent, par exemple, que « le renforcement de la discipline financière au sein du groupe peut avoir pour effet de perturber les systèmes informels de crédit et d'assurance basés sur la parenté et d'autres liens sociaux » (Attanasio *et al.*, 2015 : 114).

sociales, tant en termes de protection matérielle que d'identité. Au-delà des *randomistas*, de nombreux acteurs du monde du développement – décideurs politiques, praticiens et certains chercheurs –, considèrent la dépendance à la fois comme un problème politique (l'aide est coûteuse) et comme un problème moral (la dépendance est réputée être incompatible avec la liberté individuelle). Cependant, dans de nombreux contextes, les liens et la dépendance vis-à-vis des autres constituent à la fois un mode d'action et une stratégie délibérée. Par conséquent, l'agentivité des gens et leur « liberté de choix » se traduisent plutôt par la capacité à choisir certaines formes de dépendance et d'interdépendance[23].

En définitive, la conclusion de l'introduction générale sur l'amélioration de l'autosuffisance, ainsi que celle sur la « liberté de choix », est induite par des interprétations arbitraires des résultats économétriques (voire des extrapolations des conclusions de certaines des RCT). Ces interprétations sont sous-tendues par une conception singulière de l'autonomie et de la liberté individuelles, et donc par leur propre théorie du changement, qui voit les gens comme des atomes isolés, faisant fi des multiples rôles que jouent les interdépendances sociales à différents niveaux et jugeant implicitement néfastes ces interdépendances. Ces deux conclusions – « autosuffisance » et « liberté de choix » – ont néanmoins été incluses dans le *Policy Bulletin* (J-PAL et IPA, 2015), qui a ensuite été largement diffusé par de nombreux blogs et réseaux de discussion et considéré comme un résultat incontestable de cette recherche.

Microcrédit et surendettement

Une conclusion majeure du numéro spécial est que le microcrédit n'est pas le « piège de la dette » dénoncé par ses détracteurs. Tout d'abord, il convient de noter qu'aucune étude scientifique n'est mentionnée dans l'introduction générale, comme si le « piège de la dette » était une réalité anecdotique sans aucun fondement empirique. Il est vrai que les médias ont abondamment parlé des diverses crises de remboursement des microcrédits (tout comme ils avaient fait l'éloge du microcrédit à ses débuts). Cependant, presse mise à part, il existe un vaste corpus scientifique traitant du surendettement des ménages dans les pays du Sud et du rôle joué par le microcrédit (Schicks et Rosenberg, 2011 ; Guérin *et al.*, 2013b ; 2015 ; Schicks, 2013), y compris dans les pays couverts par le numéro spécial. Un certain nombre de problèmes se posent ici.

Le premier concerne la validité externe, où l'extrapolation se fait sans tenir compte de la singularité des contextes étudiés, ni de la dimension « marginale » des zones étudiées (Wydick, 2016). Les six RCT se sont concentrées sur des zones et des populations qui étaient censées ne pas avoir accès au microcrédit[24].

23. Pour avoir un aperçu général de la façon dont l'anthropologie aborde cette question, voir par exemple Ferguson (2015).

24. Comme mentionné ci-dessus, cette « virginité » était en fait un leurre et toutes les populations témoins avaient accès au microcrédit. Toutefois, le marché n'était pas aussi saturé qu'il aurait pu l'être ailleurs, de sorte que le risque de surendettement était moindre.

Or, par définition, le problème du surendettement y est moins aigu que dans les zones et les populations précédemment exposées au microcrédit. C'est donc une tautologie de dire qu'il n'y a pas là de « piège de la dette ». Le surendettement de certains clients du microcrédit a pourtant été documenté et parfois mesuré dans quatre des pays étudiés[25]. Le fait que les RCT ne l'aient pas quantifié ne leur permet pas de conclure que le piège de la dette n'existe pas. Contrairement à ce que suggèrent les auteurs de l'introduction générale, la littérature disponible ne se contente pas d'« anecdotes ». Les chercheurs ayant travaillé sur ce sujet démontrent (le plus souvent de manière qualitative) le rôle du microcrédit en se basant sur une analyse détaillée de ses caractéristiques spécifiques par rapport à d'autres sources d'endettement, en particulier la rigidité des conditions de remboursement et la faible tolérance aux impayés. Dans certains contextes et certaines IMF, cette tolérance zéro prend la forme de procédures coercitives de remboursement[26]. Ces chercheurs proposent également une analyse nuancée et contextualisée, mettant en évidence le rôle de l'économie globale dans le contexte mondial (notamment la stagnation et la baisse des revenus réels face à des besoins croissants), ainsi que le rôle ambivalent du microcrédit (pour certains emprunteurs, le microcrédit peut être un moyen de rembourser des dettes informelles et de réduire le surendettement)[27]. Le lien de causalité entre le microcrédit et le surendettement peut ne concerner qu'une minorité de clients du microcrédit (ce qui nous ramène à la question de l'hétérogénéité), mais ses répercussions (appauvrissement, exclusion sociale, suicide, etc.) (SCHICKS, 2013) sont suffisamment tragiques pour justifier que les *randomistas* prennent le phénomène plus au sérieux.

Le second problème est l'extrapolation des six études de cas par les auteurs de l'introduction. Même dans les zones et les populations récemment exposées au microcrédit, le surendettement ne peut être exclu. La RCT de Bosnie-Herzégovine a été conduite dans un contexte de crise de surendettement avéré, que les auteurs mentionnent comme élément de contexte (AUGSBURG *et al.*, 2015 : 185). Cette RCT conclut spécifiquement que le groupe de traitement avait des difficultés de remboursement (AUGSBURG *et al.*, 2015 : 199-201) et que celles-ci constituent un symptôme potentiel de surendettement[28]. La RCT ne permet de conclure ni à l'existence du surendettement ni au rôle du microcrédit. Cela étant, l'existence d'un « piège de la dette » ne peut être exclue.

25. Pour le Mexique, voir MORVANT-ROUX (2013), ANGULO SALAZAR (2013), HUMMEL (2013), ROZAS (2014). Pour l'Inde, voir TAYLOR (2011), GUÉRIN *et al.* (2013b), JOSEPH (2013), PRATHAP et KHAITAN (2016). Pour la Bosnie-Herzégovine, voir BATEMAN (2010), MAURER et PYTKOWSKA (2011), OPEM et GORONJA (2013). Pour la Mongolie, voir JAVOY et ROZAS (2013).

26. En Inde, par exemple, la poursuite des mauvais payeurs sur leur lieu de travail ou à leur domicile, les dénonciations et insultes publiques, la sollicitation de parents, les menaces physiques, la confiscation de biens et de documents administratifs ; dans certains cas, les plus récalcitrants ont été ligotés sur la place publique ou en plein soleil (ARUNACHALAM, 2011 ; SERVET, 2011).

27. À l'heure où nous finalisons ce chapitre (octobre 2019), les Nations unies viennent de se saisir de cette question en commandant un rapport sur le sujet, ce qui semble indiquer que le problème existe bel et bien : https://www.ohchr.org/EN/Issues/Development/IEDebt/Pages/ReportPrivateDebt.aspx.

28. Les défauts de paiement peuvent aussi être des défaillances « stratégiques » exprimant un refus de rembourser, notamment dans le contexte d'une crise d'impayés.

Dans la RCT mongole, les auteurs prennent soin de préciser que leur étude ne mesure pas le surendettement, mais uniquement les défauts de remboursement, qui sont deux choses distinctes[29]. L'introduction du numéro spécial ne fait pas référence à ces précisions.

Au Maroc, une étude qualitative menée par l'un d'entre nous en même temps que la RCT a conclu à une faible propension à l'endettement dans les zones rurales, pour des raisons culturelles et religieuses (Morvant-Roux *et al.*, 2014). Ce constat général, valable « en moyenne », n'exclut cependant pas les problèmes de surendettement dans une fraction de la population. Étant donné que le Maroc a également connu une crise de remboursement du microcrédit (que les auteurs ne mentionnent pas, bien qu'elle ait eu lieu pendant la RCT), les IMF concentrent leur offre sur une minorité de clients jugés solvables et fiables. Ces clients sont donc surexposés au microcrédit, et certains d'entre eux rencontrent effectivement des problèmes de surendettement (Morvant-Roux et Roesch, 2015).

Tout comme la Bosnie-Herzégovine, l'Inde a été frappée par d'importantes crises d'impayés dans le domaine du microcrédit : dans le district de Krishna, dans l'Andhra Pradesh, en 2006, puis dans une petite ville du Karnataka en 2009 et dans tout l'État de l'Andhra Pradesh en 2010. Les analyses de cette crise, tant quantitatives que qualitatives, ont révélé l'existence d'un problème de surendettement pour une partie des clients. Le surendettement des populations pauvres, avec ou sans microcrédit, a également été documenté en dehors des zones de crise de la dette, y compris dans les zones urbaines. Comme indiqué plus haut, la RCT indienne a été menée de 2005 à 2010 dans des quartiers périphériques d'Hyderabad nouvellement exposés au microcrédit. Mais comment est-il possible d'extrapoler à partir de cette étude de cas très spécifique alors qu'il existe un vaste ensemble de preuves démontrant l'existence du surendettement ? Sur cette question, l'article de (Banerjee *et al.*, 2015b : 23) ne cite qu'un seul article de presse : « Les anecdotes sur les entrepreneurs très prospères ou les emprunteurs très endettés ne nous disent rien sur l'effet de la microfinance sur l'emprunteur moyen, et encore moins sur l'effet de l'accès à la microfinance pour le ménage moyen ». Au vu de l'état d'alerte sur le niveau de surendettement des populations indiennes pauvres, et compte tenu de l'extrême spécificité des zones qu'ils étudient, n'est-ce pas leur propre étude qu'il faudrait peut-être qualifier d'anecdotique ?

Enfin, on peut se demander si la mesure de l'endettement des ménages a été correctement effectuée. La collecte de données fiables sur l'endettement exige un certain nombre de précautions pour les raisons suivantes : le tabou de l'endettement, exacerbé lorsque les IMF prétendent éradiquer l'emprunt informel au motif qu'il encourage les clients à dissimuler leurs dettes informelles, la diversité des terminologies employées et l'éventail des dettes susceptibles d'être détenues par différents membres de la famille sans qu'ils ne partagent nécessairement cette information. Compte tenu des approximations observées aux autres étapes de

29. Comme certaines défaillances peuvent être stratégiques, de bons taux de remboursement peuvent masquer les sacrifices consentis pour honorer des dettes, ce que reconnaissent les auteurs de la RCT en Mongolie (Attanasio *et al.*, 2015 : 114, note 25).

la collecte et de l'analyse des données (partie « Validité et portée du numéro spécial : évaluation critique »), il n'est pas déraisonnable de mettre en doute la capacité des *randomistas* à concevoir un questionnaire rendant dûment compte de l'endettement des ménages. Il convient toutefois de noter que cette difficulté n'est pas propre aux *randomistas*. La collecte de données fiables sur les revenus dans les pays du Sud a nécessité des décennies d'apprentissage pour adapter les outils statistiques aux contextes où les ménages jonglent avec différentes sources de revenus, y compris des sources informelles. Ce travail reste encore à faire sur la dette, qui demeure mal mesurée et souvent sous-estimée.

Conclusion et discussion

Compte tenu des nombreuses limites et lacunes que nous avons constatées avec la randomisation, appliquée ici au microcrédit, on peut se demander pourquoi les RCT ont connu un tel succès dans les milieux académiques, médiatiques et politiques. Nous avons déjà analysé les raisons de cette contradiction (Bédécarrats *et al.*, 2019b) en explorant l'économie politique de ce qui est devenu une véritable industrie (Ravallion, chap. 1, ce volume). Comme pour toute industrie, le marché de l'évaluation d'impact met en lien une offre et une demande. Nous avons exploré ces deux éléments en détail, en montrant que la demande est double, stimulée à la fois par la communauté des bailleurs de fonds et par le monde académique, tandis que l'offre est largement façonnée par des entrepreneurs scientifiques qui semblent avoir créé un nouveau « *business model* » en vue d'imposer un monopole et une position de rente sur le marché de l'impact de l'évaluation. La manière dont les données ont été produites et analysées, comme nous l'avons fait ici, est une illustration de cette stratégie de domination. Outre la technique de la table rase (partie « Faire table rase des recherches antérieures »), trois autres stratégies paraissent essentielles : s'affranchir d'une « culture de la donnée », ignorer la critique et contourner certaines règles de l'éthique scientifique.

S'affranchir d'une « culture de la donnée »

Les nombreuses erreurs de collecte et de saisie des données observées dans la RCT marocaine semblent suggérer un certain manque d'expérience et de connaissances, comme si les compétences purement techniques requises pour la deuxième étape (économétrie : traitement des questions de biais, sélection et identification d'un contrefactuel) dispensaient les chercheurs de l'obligation de se doter du savoir-faire requis pour la première étape (collecte de données de bonne qualité). Dans quelle mesure cette considération s'applique-t-elle aux autres RCT ? Malheureusement, cette question reste pour l'heure ouverte, car seules des réplications complètes pourraient y répondre. Cela étant, il est clair

que les *randomistas* ont tendance à ignorer les débats autour de la collecte de données (comme ils le font pour la question de l'éthique : voir Abramowicz et Szafarz, chap. 10, ce volume). Dans la plupart des protocoles de recherche empirique quantitative, il y a une véritable division du travail entre les collecteurs et les analystes de données : les premiers sont des statisticiens, les seconds sont des économistes (économétriciens ou spécialistes de thématiques spécifiques). À quelques exceptions près (DEATON, 1997 ; GROSH et GLEWWE, 2000), peu de personnes peuvent se targuer de couvrir les deux extrémités du spectre. Il s'agit de métiers à part entière, qui requièrent des compétences et une formation distinctes. Les statisticiens sont responsables de l'exactitude de la mesure, les économistes de sa pertinence, de son analyse et des relations et interactions entre les données. Ces deux activités sont indispensables pour générer des résultats « raisonnables », même si les statisticiens ont moins de prestige social que les économistes (DESROSIÈRES, 2013a). Compte tenu des compétences impliquées et du mode de fonctionnement des revues académiques, tous les efforts sont concentrés en amont sur la conception d'un processus de randomisation « intelligent » et en aval sur les estimations économétriques des impacts, le but étant de publier des articles dans des revues de premier plan.

La déconnexion entre les chercheurs et le terrain est une autre illustration de ce déni à l'égard de la culture des données. Cette déconnexion est particulièrement aiguë chez J-PAL. Son organisation hiérarchique impose une stricte division du travail entre les chefs de projet, les doctorants et les personnels de terrain (superviseurs et enquêteurs). Ces derniers se voient confier des responsabilités considérables pour lesquelles ils ne sont sans doute pas suffisamment formés (JATTEAU, 2018). Cette division du travail est une pratique fréquente dans le domaine des sciences naturelles et des sciences du vivant, mais elle n'empêche pas les chefs d'équipe de rester en contact régulier avec la chaîne de production des données, y compris pour les expériences *in vivo*. De plus, les équipes sont tenues de respecter des protocoles précis pour valider la rigueur des expérimentations réalisées. Ce n'est probablement pas le cas ici, étant donné les dizaines de RCT impliquant les personnalités les plus éminentes de la mouvance RCT (BÉDÉCARRATS *et al.*, 2019b). Cette déconnexion a été exacerbée par l'expansion exceptionnellement rapide de J-PAL, comme indiqué plus haut.

Cette croissance, conjuguée à une gouvernance extrêmement centralisée, fait que seule une poignée de chercheurs dirigent un nombre considérable d'expérimentations, ce qui fait planer le doute sur leur capacité réelle à s'impliquer sérieusement sur chaque RCT (et fait le lit de la déconnexion avec le terrain). En février 2019, Esther Duflo avait ainsi 64 RCT à son actif, soit un peu plus de quatre nouvelles RCT par an. Mais c'est Dean Karlan qui est de loin le plus prolifique, avec 100 RCT (dont 42 en cours). Dès lors, on est en droit de se demander ce qu'ils donnent réellement de leur personne pour produire chacun des résultats des RCT qu'ils signent. De fait, la signature d'un chercheur *randomista* de haute volée semble plus avoir valeur de laissez-passer pour se faire publier dans une revue de premier plan, dans le cadre d'une stratégie globale de randomisation, que de gage de qualité de la recherche.

Ignorer les critiques

Si les *randomistas* ont élaboré un discours universel sur l'impact du micro-crédit à l'issue de ce numéro spécial (et de publications ultérieures), d'autres acteurs ont pour leur part tiré des conclusions différentes de ces mêmes études (Kabeer, 2019). Là encore, la RCT marocaine en est une illustration typique. Dès 2009, alors que l'enquête finale était encore en cours, le bailleur de la RCT a commencé à partager publiquement son point de vue sur la méthode RCT en se basant sur la RCT marocaine et sur une autre étude RCT menée simultanément au Cambodge. Les conclusions étaient claires : elles mettaient en évidence les difficultés rencontrées par la méthode pour produire des évaluations d'impact rigoureuses en raison des multiples manquements au protocole que l'équipe de recherche du bailleur avait partiellement identifiés (problème de représentativité et changement de produit) et des contraintes de temps qui obligeaient à se focaliser sur des impacts de court terme. Si les conclusions de l'équipe de recherche du bailleur de fonds ont été présentées publiquement et publiées à de nombreuses reprises (Bernard *et al.*, 2012), elles sont restées lettre morte pour l'équipe de la RCT (Bédécarrats *et al.*, 2019b).

Notre propre expérience de la RCT marocaine, bien qu'illustrative, donne un bon exemple de ce que pourrait être une stratégie consistant à ignorer les critiques, jusqu'à un certain point. Au cours de nos recherches critiques dans les champ du développement, nous avons invité certains des partisans les plus virulents des RCT à engager un débat scientifique (une controverse) en de nombreuses occasions (sessions dédiées lors de conférences internationales). Nous n'avons à ce jour reçu aucune réponse. Nous avons également invité dix des plus célèbres *randomistas* à participer à cet ouvrage collectif afin d'équilibrer les points de vue sur les RCT. Ils ont tous refusé. Concernant notre examen critique de la RCT marocaine, nous avons informé les auteurs du déroulement, puis de la publication de notre réplication (Bédécarrats *et al.*, 2019a). En même temps, nous avons rédigé un *Comment* et suggéré que l'*AEJ:AE* le publie avec une *Answer to the Comment* des auteurs, comme le veut la pratique dans de nombreuses revues. L'*AEJ:AE* a refusé l'offre au motif que la revue ne publie pas de commentaires. Enfin, lorsque notre article a été évoqué dans des blogs influents et dans la presse, Crépon *et al.* (2019) ont rédigé un *Rejoinder* (51 pages) en recourant à des analyses sophistiquées pour expliquer que leurs résultats originaux étaient robustes : *double post lasso procedure*, *Benjamini-Hochberg false discovery rate correction of multiple testing*, *bayesian hierarchical model*, et analyse d'apprentissage automatique, entre autres, concluant que notre réplication n'était pas scientifique. Ils ont publié leur *Rejoinder* sur leur site internet et nous ont enjoint de le publier sur le site DIAL, ce que nous avons fait en bonne et due forme. Ils ont également informé la hiérarchie de l'AFD. L'*International Journal for Re-Views in Empirical Economics* (IREE) a suggéré que les deux parties publient une version courte du *Rejoinder* avec notre réponse (*Rebuttal of the Rebuttal* ; Bédécarrats *et al.* [2019c]). Au vu des conclusions totalement contradictoires des deux documents, nous avons proposé de demander l'avis

d'un tiers pour déterminer si nous devions retirer notre réplication (Bédécarrats *et al.*, 2019a) ou le papier initial (Crépon *et al.*, 2015) en fonction de la conclusion. Une fois de plus, ils ont décliné l'invitation. Ces épisodes illustrent deux caractéristiques de la stratégie *randomista*. Premièrement, contrairement à l'un des principaux arguments de vente des RCT (la simplicité de la méthode, par rapport à la « boîte noire » des méthodes économétriques alternatives), ce type de RCT est extraordinairement complexe. Dans leur *Rejoinder*, ils ont rajouté de la complexité à un protocole de randomisation déjà extraordinairement complexe (l'un des trois paradoxes que nous avons cherché à expliquer dans Bédécarrats *et al.* [2019b]). Deuxièmement, ils ont contourné les normes scientifiques en ne mentionnant pas leurs codes, en refusant une révision par les pairs de leur *Rejoinder* et, enfin, en éludant une controverse scientifique équitable.

Contourner l'éthique scientifique

En plus de faire fi de tout ce qui ne relève pas des RCT, les *randomistas* ont éludé certaines règles de base de la conduite scientifique. Ce problème semble gagner de l'ampleur dans l'ensemble de la communauté scientifique (Heckman et Moktan, 2018). Or, s'il n'est certes pas spécifique au J-PAL ou à la communauté des *randomistas*, il est particulièrement patent ici. Dans le monde de la recherche, la validation des connaissances est basée sur le principe de l'« examen par les pairs », c'est-à-dire un examen collectif pratiqué par des chercheurs qui jugent de manière critique et anonyme le travail de leurs homologues. Mais, pour ce faire, de nombreuses règles éthiques doivent être respectées, à commencer par la gestion des conflits d'intérêts entre les auteurs et les membres des comités de rédaction des revues. Le favoritisme éditorial est un processus reconnu et démontré, notamment chez les économistes (Fourcade *et al.*, 2015). Le numéro spécial est révélateur à cet égard. Les trois éditeurs scientifiques du numéro sont tous membres du J-PAL (Banerjee *et al.*, 2015c). Outre l'introduction générale, chaque éditeur a cosigné un article et deux d'entre eux étaient membres du comité de rédaction (Banerjee et Karlan). Esther Duflo est à la fois rédactrice en chef (et fondatrice) de la revue et co-auteure de deux des six articles. De surcroît, comme près de la moitié des auteurs des articles (11 sur 25) sont également membres du J-PAL et comme quatre autres sont des professeurs ou des doctorants affiliés au J-PAL, la revue a quelque peu dérogé aux principes d'examen par les pairs censés régir toute publication scientifique. Ce seul exemple illustre l'extraordinaire densité des liens entre les promoteurs des RCT, densité par ailleurs démontrée à une échelle beaucoup plus large par Jatteau (2016).

Que reste-t-il du numéro spécial ?

À l'issue de notre exploration, qu'avons-nous appris des RCT sur le microcrédit dans le domaine du développement ? Pour revenir au titre de ce chapitre, si le microcrédit n'est pas un miracle, comme le défend le numéro spécial, que sont

les RCT sur le microcrédit : *miracle ou mirage ?* Nous allons faire le point sur nos résultats et proposer quelques pistes de réflexion.

Nous commencerons par aborder les questions de validité interne, le point fort des RCT salué par tous. Premièrement, comme l'admettent les *randomistas* eux-mêmes, il n'y a pas de preuves solides que le microcrédit est transformateur, tout comme il n'y a pas de preuves solides qu'il ne l'est pas (Banerjee *et al.*, 2015c). Étant donné que les RCT pâtissent généralement d'un manque de puissance statistique en raison du faible taux d'adhésion et de conformité, nous ne pouvons pas nous prononcer. Deuxièmement, comme l'admettent également les *randomistas*, les effets hétérogènes sont peut-être la norme. Le microcrédit peut être transformateur pour certains et pas pour d'autres (ou pire, le microcrédit peut être négativement transformateur). Ici encore, comme les RCT pâtissent généralement d'un manque de puissance statistique en raison du faible taux d'adhésion et de conformité, nous ne pouvons pas nous prononcer. En outre, nous ne savons pas pourquoi certains peuvent tirer profit du microcrédit et d'autres pas (ou peuvent subir une pénalité « transformatrice »). Nous n'avons aucune idée des canaux par lesquels le microcrédit pourrait avoir un impact. Troisièmement, la piètre qualité des données et les erreurs de mesure peuvent conduire à reconsidérer certains des résultats pris jusque-là pour acquis. À cet égard, les nombreux problèmes que nous avons identifiés avec la RCT marocaine méritent d'être pris au sérieux. La RCT marocaine est peut-être un cas unique (le mouton noir), mais, en ce cas, ses conclusions devraient être définitivement révoquées. Cela aurait deux répercussions directes. La démonstration globale serait affaiblie. L'échantillon « assez représentatif » utilisé pour tirer des conclusions générales deviendrait « moins représentatif ». Ses bonnes propriétés mises en avant dans le numéro afin d'estimer les problèmes d'entraînement et prévoir le taux d'adhésion et ses stratégies d'échantillonnage pour traiter la question de la faible conformité et du manque de puissance statistique partiraient en fumée. Mais ce n'est peut-être pas un cas unique (bien qu'il y ait peu de chances que d'autres RCT soient aussi peu performantes), auquel cas nous sommes face à un problème structurel. La seule façon de le savoir serait de procéder à des réplications complètes, comme la nôtre. Nous préconisons vivement cette piste de recherche. Quatrièmement, nous avons montré que de nombreuses interprétations de l'impact du microcrédit, qui sous-tendent la théorie du changement, sont biaisées, tandis que certains impacts évidents (ou causes du faible taux d'adhésion) ne sont même pas pris en compte. En outre, d'autres problèmes génériques subsistent, comme les effets d'équilibre général, les politiques macroéconomiques, etc. (il s'agit dans les deux cas de problèmes de validité interne et externe).

Deuxièmement, la validité externe n'a jamais été le point fort des RCT. Notre évaluation ne change rien à cet état de fait. Les critiques habituelles, qu'il n'est pas utile de citer à nouveau ici, sont toujours d'actualité. La nouveauté du numéro spécial est qu'il examine en tandem différentes RCT sur le microcrédit. Cependant, l'accumulation de cas individuels ne résout pas le problème. Ce que l'on gagne à diversifier des contextes géographiques, mais hyper spécifiques,

se perd dans l'hétérogénéité croissante des traitements, des opérateurs, etc. Un type de produit particulier peut fonctionner dans un contexte, mais être inopérant dans un autre. Les modifications apportées aux produits et aux règles d'octroi ne fonctionnent pas de la même manière dans le « monde réel ». Enfin, les questions éthiques restent largement ignorées malgré des divergences importantes par rapport aux bonnes pratiques dans le domaine médical et même par rapport aux RCT sociales dans les pays développés.

Si l'on tient compte de tout cela, que reste-t-il ? Pour paraphraser BANERJEE et DUFLO (2011 : 167), cités par Labrousse (chapitre 8, ce volume), près de dix ans et des dizaines de RCT sur le microcrédit plus tard : « Malheureusement, […] encore aujourd'hui, il y [a] en fait très peu de preuves, dans un sens ou dans l'autre, sur ces questions. Ce que [les *randomistas*] appellent des preuves ne sont en fait que des études de cas […]. » Bien que l'on ne sache pas précisément ce qu'il reste à ce stade, il est en tout cas une chose qui s'est évanouie : l'énorme quantité d'argent et de ressources dépensées, dont une partie retirée à des alternatives et à d'autres usages. Cela vaut-il la peine de dépenser des millions de dollars pour ne publier qu'un seul article académique pour chaque RCT (tabl. 5) ? Ne serait-il pas plus utile d'affecter ces fonds au financement du système statistique public d'un pays en développement afin de collecter une quantité massive de données d'observation représentatives sur le long terme ? Si les promoteurs des RCT ont admis certaines des lacunes méthodologiques évoquées dans ce chapitre, leur réponse pour les résoudre est « Nous voulons plus de RCT ! » Or, comme les RCT n'ont pas tenu leurs promesses, du moins les promesses faites au monde entier par les *randomistas* au cours de ces deux dernières décennies, il serait tout aussi légitime de dire : « Nous ne voulons plus de RCT ! »

Cette proposition peut sembler abusive, mais le raz-de-marée randomiste a été si puissant (comme le montre la façon dont ils ont balayé le passé en ignorant toutes les études non expérimentales) qu'une petite poussée dans l'autre sens ne ferait pas de mal pour rééquilibrer les choses. Notre objectif n'est toutefois pas de discréditer la méthode RCT, mais de reconnaître sa véritable valeur en récusant le piédestal sur lequel elle trône actuellement. Plutôt que « Nous voulons plus de RCT ! », nous préconisons ce message : « Nous ne voulons plus de RCT autonomes ! » Si les RCT peuvent tout à fait rester appropriées et légitimes pour certaines politiques précisément circonscrites, elles doivent néanmoins être menées selon les règles. De plus, elles ne se suffisent jamais à elles-mêmes. Il est à la fois nécessaire et possible d'utiliser d'autres méthodes sans compromettre la rigueur scientifique. Comme nous l'avons vu ici, ce pluralisme devrait être une exigence, notamment pour compléter les RCT en les contextualisant, tant avant la collecte des données que pour leur analyse. Le pluralisme est aussi une exigence pour l'ensemble des questions, projets et politiques de développement qui ne se prêtent pas aux RCT, et le microcrédit, avec ses interventions relativement bien ciblées, en est un bon exemple au vu du faible taux d'adhésion et de la complexité de ses effets. Malheureusement, pour de nombreux promoteurs des RCT, et le J-PAL en particulier, « les RCT

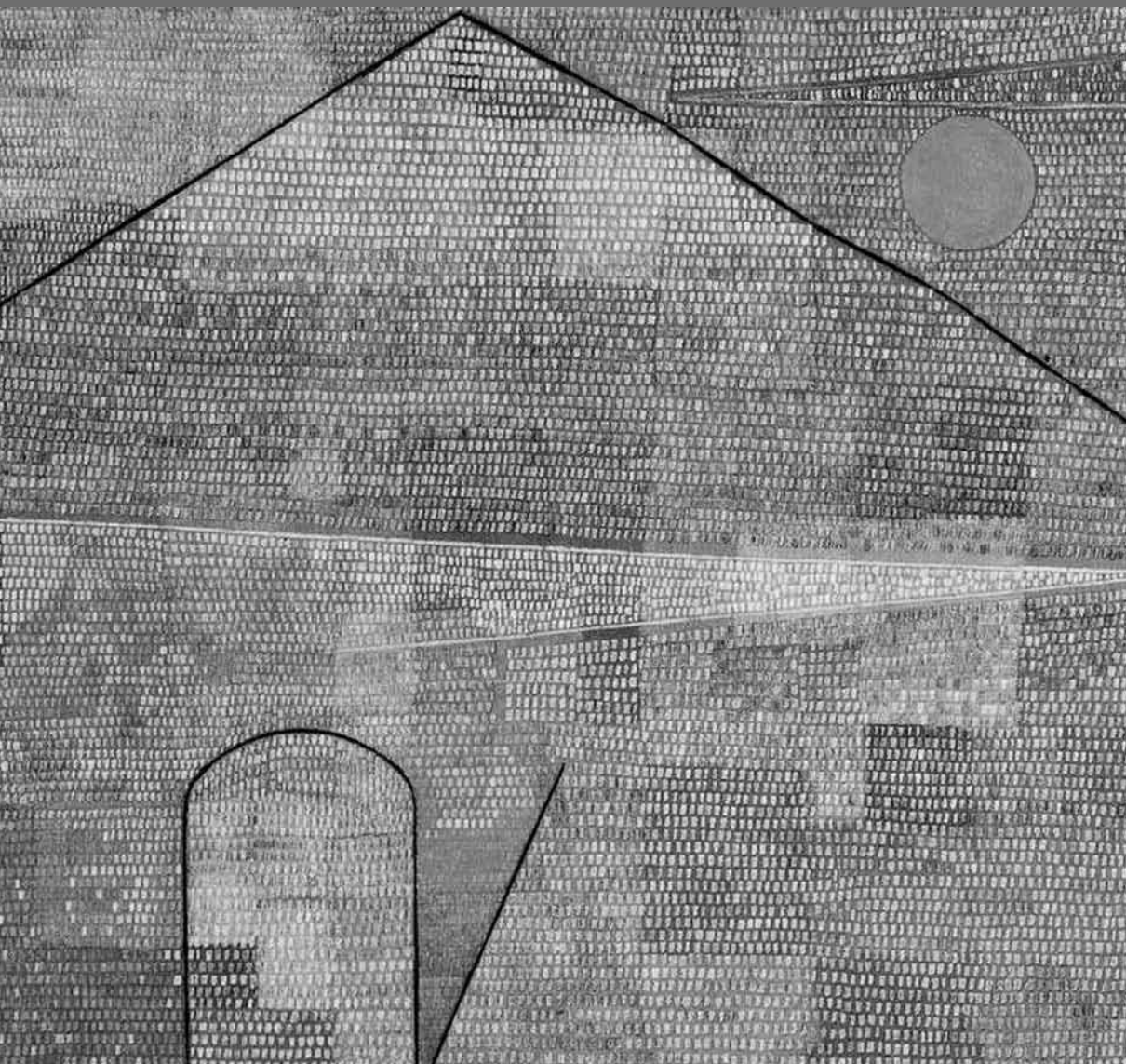
ne sont pas seulement en tête de liste des méthodes approuvées, pour eux, il n'y a rien d'autre sur la liste », (Ravallion, chap. 1, ce volume).

Remerciements

Nous remercions les participants de l'atelier de mars 2019, qui a rassemblé la plupart des contributeurs à l'ouvrage, ainsi que Solène Morvant-Roux, Jonathan Morduch et Martin Ravallion pour leurs commentaires sur une version antérieure du chapitre.

Partie 3

Économies politiques



Chapitre 8

La supériorité rhétorique de *Poor Economics*

Agnès LABROUSSE

Poor Economics : un succès frappant, une rhétorique persuasive

Comme l'a montré MCCLOSKEY (1983), la rhétorique est omniprésente en économie. Cependant, cette dimension discursive est souvent occultée, car parmi les économistes domine un credo d'objectivité scientifique associé à la rigueur du chiffre. Le livre *Poor Economics* du J-PAL ne fait pas exception à cette règle : sa revendication de légitimité repose sur des chiffres concrets produits par des évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT), où l'utilisation de preuves anecdotiques est explicitement prohibée. Par rapport à d'autres courants économiques *mainstream*, ce discours épouse des caractéristiques rhétoriques particulières. Comme nous le verrons, ce sont précisément ces caractéristiques qui le rendent persuasif et qui contribuent au succès de ce laboratoire et de la randomisation auprès de divers publics. Ce formidable succès n'en demeure pas moins déroutant. En dépit de limites importantes et bien identifiées en médecine (LABROUSSE, 2010), en économie (BÉDÉCARRATS *et al.*, 2019b ; Pritchett, chap. 2, ce volume ; Ravallion, chap. 1, ce volume) et dans le champ de l'évaluation (BERNARD *et al.*, 2012 ; Picciotto, chap. 9, ce volume), le succès des RCT ne faiblit pas et la « bulle » n'a pas encore éclaté. Nous soutenons ici que la rhétorique ingénieuse du J-PAL constitue une pièce maîtresse de cette énigme.

Nous allons nous pencher ici sur le livre à succès *Poor Economics* (*Repenser la pauvreté*) de BANERJEE et DUFLO (2011 : IX)[1]. Cet ouvrage condense la rhétorique du Abdul Latif Jameel Poverty Action Lab (J-PAL) – le plus grand laboratoire au

1. Contrairement à *Poor Economics*, leur dernier livre, *Good Economics for Hard Times* (BANERJEE et DUFLO, 2019), ouvre un horizon plus large et ne s'appuie donc pas exclusivement sur les RCT : il fait référence à un grand nombre de publications et de sujets qui ne relèvent pas des travaux du J-PAL.

monde à travailler sur la pauvreté à l'aide des RCT – et sa prétention à « lutter contre la pauvreté avec des chiffres rigoureux [*hard numbers*] ». Il est donc représentatif de l'usage dominant des RCT en économie[2] et de sa justification. Le livre s'adresse à un large public, dépassant la simple sphère académique pour toucher les membres des agences internationales et gouvernementales, les ONG, ainsi que les journalistes, les étudiants et les citoyens. Il condense les résultats de plusieurs centaines d'expérimentations sur une multitude de sujets. Le sujet de l'ouvrage est « *how to fight global poverty* » et « *how the poor really live their lives* » (p. 14)[3]. *Poor Economics* a fait l'objet d'une large couverture académique et médiatique. Il a été salué par des économistes de renom aussi divers que Robert Solow, Amartya Sen, William Easterly ou Anne Krueger, mais aussi par des philanthropes comme Bill Gates, des journaux généralistes de qualité (*The New York Times*, *The Guardian*, etc.), ainsi que des revues économiques spécialisées : *The Wall Street Journal* a fait l'éloge du livre, qui s'est vu décerner le prix « *Business Book of the Year* » par le *Financial Times* et Goldman Sachs. Il a suscité un nombre impressionnant de comptes rendus et de citations (2 713, selon *Publish or Perish* le 4 juin 2019). Aux États-Unis, le livre a été publié par *Public Affairs*. Comme indiqué p. 297, cette maison d'édition a également publié « Gandhi, Nasser, Toynbee, Truman et quelque 1 500 autres auteurs ».

À la lecture de *Poor Economics*, il apparaît que le livre est certes truffé de chiffres, mais aussi, ce qui est plus surprenant, d'anecdotes. De fait, l'efficacité argumentative de l'ouvrage passe par diverses formes de mise en récit, plus que par le seul pouvoir des chiffres nus. L'objectif de cette analyse est (1) d'établir, par une analyse textuelle, la présence et la raison d'être de processus rhétoriques saillants ; (2) d'analyser leurs effets de persuasion et de connaissance ; (3) d'éclairer à la fois le succès des RCT et certaines limites de cette technique à la mode – notamment les questions auxquelles elle nous rend aveugles.

Cadre théorique : rhétorique ordinaire, communautés épistémiques et analyse de discours

Pour ce faire, cette analyse puise dans le champ ouvert en économie par l'article fondateur de McCloskey (1983). Centrée sur la « rhétorique ordinaire » (*workaday rhetoric*) des économistes, cette approche intègre les arguments statistiques et les modèles formels à l'analyse des processus rhétoriques. À la différence de McCloskey, la rhétorique n'est pas envisagée comme une conversation autodisciplinée par une éthique de la discussion, pratiquée par une élite d'honnêtes gens, où les meilleurs discours – les plus persuasifs – triompheraient spontanément sur un « marché des idées économiques » exempt de toute forme de domination (Mäki, 1995). Point de marché des idées ici, mais un

2. Le J-PAL domine le champ (Jatteau, 2016). Au 5 juin 2019, le J-PAL avait conduit 952 RCT, alors que le nombre total de RCT enregistrées en sciences sociales s'élevait à 2 552, soit 37,2 % (https://www.socialscienceregistry.org/). Il faut savoir que d'autres usages des RCT existent en sciences sociales et médicales (Labrousse, 2016).

3. Les indications de pages entre parenthèses renvoient à Banerjee et Duflo (2011).

champ de lutte et de coopération entre différentes communautés épistémiques, un champ hiérarchisé et institué de production et d'évaluation des connaissances (Bourdieu, 1975 ; Chavance et Labrousse, 2018). Ces communautés épistémiques sont également des communautés discursives, cimentées par des convictions culturelles et épistémologiques partagées (Beacco et Moirand, 1995). Leur discours renvoie à d'autres discours tenus par d'autres communautés : ils sont conçus « comme une attaque, une défense, une critique ou une contribution à une position ou à un système de pensée particulier » (Skinner, 2003 : 100).

Cette analyse prend appui sur *La Rhétorique* d'Aristote. Ce philosophe définit la rhétorique comme « la faculté de considérer […] ce qui peut être propre à persuader » (livre I, chap. 2). Les notions aristotéliciennes fondamentales de *logos* (processus argumentatif utilisant la raison) et d'*ethos* (« sous quel jour apparaît l'orateur » (livre II, chap. 1) : les qualités dont l'orateur fait preuve par son discours) constituent les principes d'organisation fondamentaux de cette analyse. Elle examine également la « disposition textuelle […] conçue comme l'ordonnancement de stratégies persuasives au sein du texte (tactique d'agencement textuel) […] [et les] stratégies sur le plan de l'expression du texte (typo-disposition, distribution, typographie, ponctuation, etc.). » (Duteil-Mougel, 2005 : 3). Il s'agit de « dispositifs d'évaluation » : ils fournissent des indications sur ce que l'auteur entend mettre en lumière (Strassman et Polanyi, 1995).

Ici, les effets épistémiques d'un dispositif discursif sont tout aussi importants que ses effets persuasifs : de quelle manière ces procédés argumentatifs structurent-ils et façonnent-ils les connaissances, comment permettent-ils de mettre en lumière certains phénomènes et d'en occulter d'autres ? C'est un aspect important de l'analyse de discours : « les discours apparaissent comme des façons particulières de construire (représenter, interpréter) des aspects particuliers du processus social qui deviennent relativement récurrents et pérennes et qui, nécessairement, simplifient et condensent des réalités complexes, en incluent certains aspects tout en [en] occultant d'autres, et se focalisent sur certains aspects pour en marginaliser d'autres » (Chouliaraki et Fairclough, 2010 : 1215). Ces processus de focalisation, de réduction, de marginalisation et de mise hors champ méritent une attention particulière.

Méthodologie et plan

Dans cette perspective, j'ai commencé par une première lecture linéaire du livre. Elle m'a permis d'identifier les procédés rhétoriques saillants relatifs aux contenus chiffrés, graphiques et textuels de *Poor Economics*. Après cette phase d'exploration abductive, j'ai procédé au comptage et à l'inventaire inductif des occurrences (et, dans certains cas, des co-occurrences[4]) de ces éléments. Ce comptage permet d'examiner systématiquement l'importance relative des termes et d'observer leurs variations de forme et de contenu en contexte.

J'ai ensuite testé la présence d'autres lemmes dans le livre pour mettre à l'épreuve de manière déductive la solidité des premiers résultats de l'analyse, de manière

4. Lorsque l'inventaire a révélé des associations répétées, je les ai quantifiées.

à mettre en exergue le hors-discours (les termes absents ou rares). Si un terme est très présent, il y a de fortes chances qu'un terme apparenté soit lui aussi très présent ; inversement, un terme antagonique sera probablement peu présent. J'ai également comparé certains aspects de *Poor Economics* à d'autres ouvrages de vulgarisation scientifique rédigés par des économistes du développement (EASTERLY, 2001 ; SACHS, 2005 ; STIGLITZ, 2006) afin de mettre en évidence les spécificités rhétoriques de l'ouvrage.

L'analyse de la rhétorique dans *Poor Economics* commence par celle de la mobilisation des chiffres, suivie de l'étude de l'utilisation des graphiques, puis de l'examen de la présence et des multiples fonctions rhétoriques des anecdotes du livre, et se termine par l'identification de deux schèmes rhétoriques transversaux particulièrement efficaces.

Hard numbers : la rhétorique du chiffre, le chiffre comme figure rhétorique

Les chiffres occupent une place de choix dans *Poor Economics*. Très présents, ils sont accompagnés d'une rhétorique du nombre probant. Certains chiffres sont mis en scène comme arbitre des controverses économiques et pour représenter la vie des pauvres.

Quantifier et disqualifier

Poor Economics développe une rhétorique de la preuve par le chiffre que l'on retrouve dans toutes les productions du J-PAL. Ici, le champ sémantique de la preuve expérimentale et du chiffre est fortement présent. Le livre totalise 130 occurrences du terme *evidence* (dont 4 de *proof*), 102 de *fact(s)* (+ 31 de *in fact*), 85 de *number**[5], 84 de *experience**, 72 de *random**, 57 de *control**, 45 de *data*, 18 de *trial** et 9 de RCT. À titre de comparaison, parce qu'ils sont directement liés à son *leitmotiv*, les termes les plus significatifs du livre ont trait à la vie des pauvres, avec 584 occurrences de *poor**, 207 pour les termes liés à la santé (*health*)* et 150 occurrences cumulées pour *life** et *lives** (y compris *lifetime* ou *livestock*).

Les chiffres eux-mêmes sont omniprésents : le texte compte un total de 3 607 chiffres[6]. Parmi ces 3 607 occurrences figure un total de 236 dates, très concentrées sur la période précédant la publication (1990-2011). Rapportée au

5. Les astérisques signifient qu'il s'agit de la suite de caractères indiquée, suivie de n'importe quels caractères ou d'aucun caractère.

6. Ce décompte exclut la table des matières, les notes de fin, les remerciements, la présentation des auteurs et de l'éditeur, ainsi que les numéros de page, de section et de chapitre et les références à ceux-ci.

nombre de pages du livre, la moyenne est de 12,9 chiffres par page (dates incluses) et de 12,0 chiffres par page (dates exclues). Les randomistes semblent obéir au précepte de Kelvin : « Quand vous ne pouvez pas l'exprimer en chiffres, votre connaissance est maigre et insatisfaisante », suivant le credo scientifique des économistes (McCloskey, 1983 : 484). Il y a une moyenne de douze chiffres par page, bien que le livre ne contienne aucun tableau statistique et que six graphiques sur sept livrent une représentation théorique qui n'est basée sur aucune donnée statistique (voir la section suivante). Les chiffres sont donc mobilisés dans un récit « littéraire », une caractéristique pour le moins inattendue, nous y reviendrons.

Dans la rhétorique du J-PAL, seuls les chiffres issus de RCT sont probants et les RCT sont qualifiées de « *new powerfool tool* » (p. 21). « *The studies we use have in common a high level of scientific rigor, openness to accepting the verdict of the data* […] » (p. 23). Leur méthodologie apparaît d'autant plus rigoureuse qu'elle se calque sur la médecine : « *the cleanest way to answer such questions is to mimic the randomized trials that are used in medicine to evaluate the effectiveness of new drugs* » (p. 15). Ces chiffres seraient synonymes d'objectivité. Le reste ne serait qu'idéologie et ignorance, autant de fléaux que les « chiffres rigoureux » (*hard numbers*) des RCT permettraient de combattre. Avec l'inertie, ce sont les « trois i » auxquels il faudrait remédier pour aider efficacement les pauvres. Les auteurs en font expressément le message central de l'ouvrage :

> « *The message of this book* […] [*is that*] *ideology, ignorance, and inertia—the three I's—on the part of the expert, the aid worker, or the local policy maker, often explain why policies fail and why aid does not have the effect it should* » (p. 16).

Facile à retenir, la formule « 3I's *problem* » apparaît à cinq reprises dans le livre[7]. Les RCT sont placées au sommet de la hiérarchie des preuves et les autres méthodes sont, de fait, disqualifiées comme idéologiques et non conclusives. Dans de précédentes productions du J-PAL, les RCT étaient présentées comme le « *gold standard* » (aucune occurrence dans le livre), révélant un « prosélytisme méthodologique » (Jatteau, 2016). Ici, les RCT sont simplement le moyen le plus propre de procéder. En 2007, Banerjee avait qualifié les régressions internationales et les études de cas de « preuves sans consistance » (Labrousse, 2010). On retrouve cette idée de manière plus effacée dans *Poor Economics*, notamment dans les développements inauguraux qui critiquent les régressions macroéconomiques (p. 3-5), dont les résultats sont présentés comme incertains et censés relever de « *big philosophical questions* » (p. 4), de « *speculating on the grand scale* » (p. 5) et, dans le paragraphe suivant sur la microfinance :

> « *Unfortunately,* […] *until very recently, there was in fact very little evidence, either way, on these questions. What CGAP* [Consultative Group to Assist the Poor] *calls evidence turns out to be case studies* […] » (p. 167).

7. Ces 3 I (intérêts, institutions et idées) sont également une *contre-formule*. Ils visent à se substituer aux 3 I de leurs adversaires de l'économie politique institutionnelle *mainstream* (Hall, 1997).

Dans quelques passages, les chiffres semblent parler d'eux-mêmes : « *The data squarely rejected this view* » (p. 124), « *The data seems to squarely hand victory to the demand wallahs* » (p. 112), « […] *accepting the verdict of the data* » (p. 16). On peut trouver une variante avec « *evidence* » : « *Whose story—the activists' or the skeptics'—does the evidence support?* » (p. 44). « *The evidence suggests the opposite*. » (p. 50), « *Our evidence shows* » (p. 171). Cette figure de style faisant parler les données est une « hypostase » : une entité fictive (les données) est considérée comme un sujet actif (Lalande, 1902-1923). Le rôle décisif du randomiste dans la construction expérimentale et dans l'interprétation de ces résultats est alors occulté. Cela renvoie au point de vue plus général d'Esther Duflo : « Les évaluations sont rigoureuses. Elles ne laissent aucune place à l'interprétation. Si ça ne marche pas, ça ne marche pas. La seule chose qui reste à faire, dès lors, est d'essayer autre chose » (Labrousse, 2016 : 289-91).

99 cents, synecdoque pour la vie des pauvres

Dans *Poor Economics*, un chiffre est placé de manière récurrente sur le devant de la scène : *99 cents*. Représentant le seuil international de pauvreté absolue, ce nombre apparaît dix-huit fois dans le livre, tandis que son équivalent technique, *poverty line*, apparaît six fois (dont quatre fois dans une note explicative, p. 277). C'est une synecdoque pour décrire la vie des pauvres, en co-occurrence systématique avec *live*/*living* (« *living on less than 99 cents a day/per day* »). La première version du site internet dédié à *Poor Economics*, mentionnée à trois reprises, s'intitulait à l'origine « www.99centsthebook.com », un indice supplémentaire de l'importance de ce nombre. Il représente la vie des pauvres, sujet du livre, et devient une métonymie du livre lui-même.

Pourquoi *99 cents* ? Banerjee et Duflo (2011 : 277, note de fin) justifient ce choix en se référant aux travaux de référence de Deaton et Dupriez (2011) (données recueillies en 2005 pour le programme de comparaison internationale de la Banque mondiale). Le seuil monétaire de pauvreté absolue est fixé à 16 roupies indiennes, sur la base d'un panier de biens consommés par des pauvres, soit 99 cents en parité du pouvoir d'achat (PPA) avec les États-Unis, ajusté en fonction des indices de prix. Néanmoins, d'autres chiffres étaient disponibles. Si, en 1985, Ravallion avait popularisé le chiffre de 1,02 $ en PPA (à l'origine de la fameuse formule du « 1 dollar par jour »), en 2008, la banque mondiale avait réévalué ce seuil de pauvreté à 1,25 $ en PPA de 2005… Ces seuils sont source de controverses (Reddy et Lahoty, 2016) et, comme toutes les données, reposent sur des conventions sociales (Porter, 1995 ; Desrosières, 1998).

Ce chiffre fait appel à la fois au *logos* (l'argument « technique » précédent) et au *pathos*. Percutant, il rend palpable la situation des pauvres pour le lecteur, qui est nécessairement riche (il a pu acheter la version numérique du livre pour 9,99 $). Dans les pays du Nord, 99 cents est présenté comme un « prix symbolique », un rien du tout pour acheter un petit quelque chose. Dans les pays du

Sud, 99 cents, c'est tout ce dont les très pauvres disposent au maximum chaque jour. Pour le lecteur américain, c'est un nombre facilement mémorisable et emblématique de la société de consommation, dont les pauvres sont exclus : ce chiffre palindrome fait écho au chiffre après la virgule des prix de supermarché et au nom d'une chaîne discount (« 99 cents stores »).

Cette mise en avant de « *hard numbers* » permettant de sortir des alternatives stériles, de combattre les idéologies et d'arriver à des vérités purement objectives participe d'une vision dépolitisée du protocole expérimental et de l'aide internationale (LABROUSSE, 2016). Cette « foi dans les chiffres » (PORTER, 1995) est particulièrement prégnante dans l'économie *mainstream*. Elle fait écho à la façon dont les organisations internationales et les experts en développement (FERGUSON, 1990) promeuvent les « techniques de recherche de consensus, qui neutralisent les oppositions et les conflits et esquivent les relations de pouvoir » (HIBOU, 2011 : 136).

Représentations graphiques : récit incarné et métaphorique, cadrage cognitif

Les graphiques – sept au total – sont relativement peu nombreux dans *Poor Economics*. Voilà qui est surprenant, eu égard à la rhétorique des chiffres et à l'omniprésence des graphiques dans les médias contemporains (KOETSENRUIJTER, 2017). Cependant, les graphiques du livre ne présentent pas, à une exception près, de statistiques sur la vie des pauvres, ni de résultats de RCT : il s'agit plutôt de représentations économiques abstraites. Témoignant du sérieux des auteurs (*ethos*), ces formalisations rebuteraient probablement le lecteur non initié si elles n'étaient pas accompagnées d'une narration attrayante mettant en scène des personnages réels.

Qu'est-ce que le monde de Kennedy ? Représenter et réduire le champ des possibilités à deux diagrammes

S'ils sont utilisés avec parcimonie, plusieurs de ces graphiques jouent un rôle marquant. Ici, les deux premiers graphiques (p. 12-13) illustrent le thème principal et la colonne vertébrale du livre : « y a-t-il des pièges de la pauvreté ou non ? » L'importance de la notion de piège à pauvreté (*poverty trap*) est marquée par ses 46 occurrences dans le corps du texte. C'est considérable pour un terme spécialisé : le terme plus générique de *development* n'apparaît que 28 fois. Ces graphiques révèlent deux visions du monde : le monde selon Sachs (qui croit en l'existence des pièges à pauvreté et voit donc le monde selon la fig. 1) et le monde d'Easterly (« selon Easterly, il n'existe pas de pièges de la pauvreté »), représenté par la fig. 2.

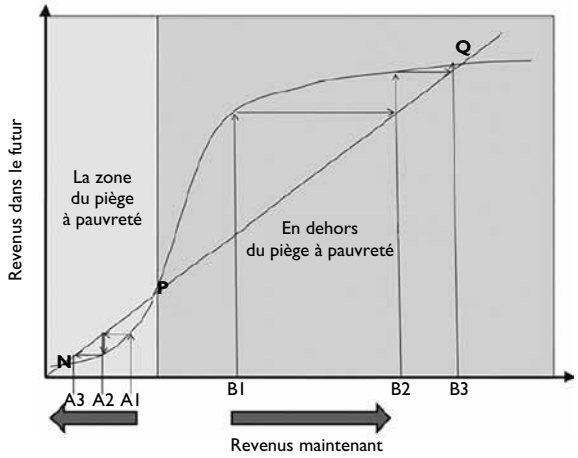


Figure 1
La courbe en S et le piège de la pauvreté.
Source : BANERJEE, DUFLO (2011).

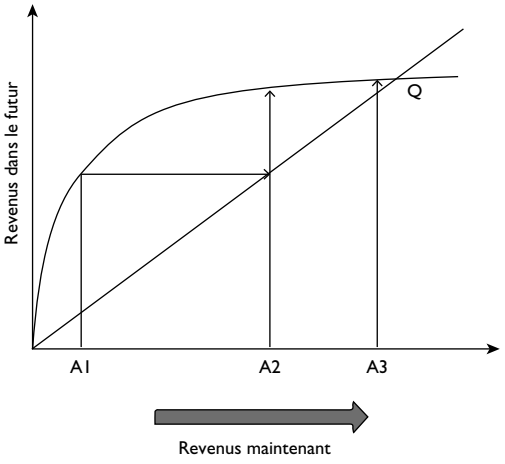


Figure 2
La forme en L inversé : pas de piège à pauvreté.
Source : BANERJEE, DUFLO (2011).

Les deux diagrammes sont accompagnés d'un commentaire qui fait de la croyance ou non dans les pièges à pauvreté une question de foi. « Pour ceux qui croient aux pièges à pauvreté, le monde ressemble à la figure [fig. 1] » ; « De nombreux économistes (une majorité peut-être) pensent cependant que le monde ressemble à la figure [fig. 2]. » Cette question est associée à un personnage réel mentionné dans le livre, Kennedy : « *So which of these diagrams best represents the world of Kennedy, the young Kenyan farmer?* » (p. 12). Cette question ne peut être tranchée que par la preuve empirique (et non théologique) issue des RCT :

> « *we will find them in some areas, but not in others*. […] We will see many instances in the chapters that follow where the wrong policy was chosen, not out of bad intentions or corruption, but simply because the *policy makers had the wrong model of the world in mind*: They *thought there was a poverty trap somewhere and there was not, or they were ignoring another one that was right in front of them* » (p. 15, souligné par l'auteure).

Dans la typologie de BRETON (1999), cette mise en récit graphique entre dans la catégorie des *énoncés de cadrage* et, plus précisément, des énoncés de cadrage manipulatoire. En effet, il repose sur un « cadrage par fausse alternative ». Les deux mondes alternatifs présentés restreignent considérablement les mondes possibles et fonctionnent comme des œillères. Ils laissent hors champ, *sans le préciser* – c'est en cela que le cadrage est manipulatoire –, des questions fondamentales en économie du développement.

Ce hors-champ est un hors discours. L'organisation de la production et de l'entreprise, les dynamiques d'innovation, les questions méso-économiques et territoriales, les flux financiers et de marchandises, locaux et internationaux, les dynamiques et les politiques macro-économiques, l'environnement et les inégalités font figure de grands absents. Ainsi, on ne trouve dans le corps du texte aucune occurrence des termes *inequal**et *unequal*, *Gini coefficient*, *income/wage disparity/ies*, *justice*, *ethics**, *dependency*, *terms of trade*, *import*, *comparative advantage*, *commodity/ies*, *stabilization*, *specialization*, *international relation**, *industrial revolution*, *capitalism*, *market economy*, *modernization*, *westernization*, *globalization*, *tariff**, *reserves*, *foreign investment*, *capital flow*/flight**, *brain-drain*, *volatility*, *instability*, *speculation/tive*, *deregulation*, *Dutch disease*, *monetary policy*, *fiscal/budgetary policy*, *redistribution**, *protectionist**, *lost decade*, *(Post) Washington consensus*, *IMF*, *structural adjustment*, *foreign debt*, *foreign investment*, *fair/free trade*, *regional development*, *value chain*, *production network*, *corporate governance/interests*, *innovation fund*, *technology gap*, *patent*, *license*, *intellectual property*, *agrarian reform*, *land grabbing*, *deforestation*, *commons/common pool*, *natural resources*, *climate change*, *greenhouse (gas)*, *biodiversity*, *public good*. Le terme *industrial policy* n'apparaît qu'une fois, tout comme *domination*, *dynamics (familial)*, *inequity* (intrafamiliale), *commerce* (l'idée de *trade credit*), *remittance*, *diversification* (des risques), *pollution* (« *pollution inspectors* »), *externalités* (« *treatment externalities* »), *global warming*, *carbon emission*, *liberalization* (« *early years of Chinese liberalization* »), *privatization* (« *privatization voucher* » pour les frais de scolarité) ou *recession*. Le terme *energy* n'est utilisé que dans son acception psychologique (3 occurrences) ; il en va de même pour 5 occurrences de *depression* sur 7. Les résultats sont similaires pour les termes *structure* et *macro* (voir section « Deux schémas rhétoriques aux puissants effets épistémiques et persuasifs »). Voilà qui est révélateur de la difficulté fondamentale des RCT à aborder les dynamiques historiques (y compris les dynamiques microéconomiques), ainsi que les questions méso et macro. Ces sujets ne se prêtent pas aux RCT.

Ces omissions deviennent patentes lorsque l'on compare ce livre à d'autres textes : en premier lieu, aux textes des fondateurs de l'économie du développement (MEIER et SEERS, 1984) ; ensuite, aux rapports d'institutions internationales qui ont rythmé les débats sur la pauvreté mondiale depuis les années 1980 : rapport de l'Unicef (CORNIA *et al.*, 1987) sur les dommages humains des programmes d'ajustement structurel (PAS), rapport du Programme des Nations unies pour le développement (PNUD, 1999) évoquant les « décennies perdues » en Afrique et en Amérique latine, rapport de la World Bank (2005) sur les enseignements tirés des échecs des PAS et des thérapies de choc ; enfin, aux écrits d'économistes établis sur le « post-consensus de Washington » (STIGLITZ, 2004 ; RODRICK, 2008). Notons que les mots-clés susmentionnés, à l'exception des termes *land grabbing*, *land reform*, *production network* et *dependency*, figurent tous dans l'ouvrage de vulgarisation scientifique *Making Globalization Work* de STIGLITZ (2006). Ces « blancs » des RCT affleurent également lorsqu'on les compare à la littérature économique consacrée sur les inégalités – de Bourguignon à Piketty[8] – ou à l'économie politique hétérodoxe. Néanmoins, le rapport *Mind, Society and Behavior* de la Banque mondiale pour 2015 rejoint, dans les grandes lignes, le point de vue de Banerjee et Duflo (World Bank, 2015). N'est considérée comme rigoureuse que l'échelle microéconomique, le méso-économique est ignoré et il est fait du macroéconomique un champ de spéculation quasi métaphysique.

La métaphore filée de la courbe en S : quand Ibu Tina tomba dans le piège à pauvreté

Tous les graphiques présents dans *Poor Economics* (BANERJEE, DUFLO, 2011) sont construits autour de la question suivante : la courbe en S du piège à pauvreté existe-t-elle ou non ? C'est notamment le cas du seul graphique basé sur des données statistiques (« *Wealth in 1999 and 2005 in Thailand* », p. 201). Rien n'est dit sur les raisons du choix de la Thaïlande et de cet intervalle de temps particulier (1999-2005), alors même que le propos est très général. Contrairement aux autres graphiques, il est peu explicite et peu explicité par les auteurs. C'est en quelque sorte un « graphique *ex machina* », dont l'intervention soudaine et mystérieuse a l'avantage de révéler une courbe en S, l'alignant sur les phénomènes du monde réel. Il est introduit avec un *do* emphatique et une pointe d'humour (la torture du S) :

> « We do see this S-shape between net worth today and net worth in the future in the real world. [*The graph*] *plots the relationship between resources the households had in 1999 and what they had five years later in Thailand.* The curve has a flat, elongated S-shape *(admittedly, we are torturing the S a little bit).* […] *What is more distinctive is the way in which the relation is fairly flat at very low levels of resources but then*

8. Piketty, qui a joué un rôle important dans l'arrivée de Duflo au MIT, est cité dans les remerciements, mais pas dans le reste du livre.

> *turns up sharply before flattening off.* This S-shape, as we saw before, generates a poverty trap » (p. 200, souligné par l'auteure).

Un autre graphique clé, « *The impact of shock on Ibu Tina's wealth* », présente une courbe en S. Il apparaît plus haut dans le livre (p. 139). Il prépare et rend tangible ce récit très général autour de la courbe en S. Ce graphique représente le basculement dans une trappe à pauvreté d'un personnage réel, Ibu Tina. Il est précédé du récit détaillé (393 mots) de la vie de cette dernière. Il narre un épisode tragique à l'origine d'un renversement de situation : un vol précipite Ibu Tina et sa famille dans la grande pauvreté, dont elle ne parviendra pas à sortir. Ce texte est suivi de sa représentation graphique : « *In Figure* [1], *we have plotted the relationship between income today and income in the future for Ibu Tina, the Indonesian businesswoman* » (p. 139).

Il est ensuite traduit sur le plan conceptuel dans un commentaire. Il fait appel à un vocabulaire économique plus prononcé et donne lieu à une morale économique (dernière phrase de l'extrait suivant) :

> « *Before the debacle of the bounced check, Ibu Tina and her husband were outside the poverty-trap zone. If we follow their path over time, we see that they were on the trajectory to eventually arrive at a decent income. But the theft wiped out all their assets. This had the effect of moving them to the poverty-trap zone. Thereafter, they made so little money that they kept getting poorer over time* […] *When the relationship between income today and income tomorrow is S-shaped, a family can plunge from being on a path to middle class to being permanently poor* » (p. 139).

Le graphique permet de *visualiser* comment Ibu Tina tombe littéralement dans une trappe à pauvreté, dont elle va rester durablement prisonnière : cette chute change le cours heureux de son destin vers un revenu décent. C'est une modalité originale de narration graphique combinant *logos* (ici le caractère abstrait du graphique) avec une incarnation individuée (Ibu Tina) et un évènement singulier (le vol) relevant du pathos, dans le cadre d'une mise en récit expressive à la fois graphique et verbale.

Ces courbes en S constituent une métaphore structurante et récurrente (pas moins de 30 occurrences de *S-Shape*). Elle met en image les trappes à pauvreté et se trouve reprise dans les autres graphiques. Cette métaphore est présentée comme une réalité vécue ou une croyance partagée et agissante des acteurs économiques, comme les enseignants, les parents d'élèves ou ce boutiquier de Gulbarga rencontré par les auteurs :

> « *As we saw,* the belief in the S-shape curve leads people to give up. *If the teachers and the parents do not believe that the child* can cross the hump and get into the steep-part of the S-curve, *they may as well not try: The teacher ignores the children who have fallen behind and the parent stops taking interest in their education* » (p. 91, souligné par l'auteure).

> « […] *once a micro-entrepreneur* realizes *that she is probably* stuck in the low part of the S-curve *and will never be able to make that much money, it may* [*be*] *difficult for her to be fully committed to her business. Imagine an entrepreneur who is below point M in Figure* [*3*]. *It could be the shopkeeper we met in Gulbarga* » (p. 223, souligné par l'auteure).

Cette courbe en S, sortie de la tête des économistes, ferait ainsi partie du monde vécu des pauvres et non des schèmes interprétatifs des auteurs. Elle est d'autant plus expressive que la métaphore est filée : dans le parcours d'obstacle scolaire, l'élève pauvre craint de franchir la bosse (*cross the hump*) lui permettant de gagner la partie abrupte de la courbe (*steep-part*) et reste en arrière, en situation de retard scolaire (double sens de *fall behind* en anglais) ; la microentrepreneuse comprend, elle, qu'elle est coincée (*stuck*) dans la partie basse de la courbe en S. Le bas de la courbe est le royaume des pauvres.

> « The shape of the curve is key: *It is very flat at the beginning, and then rises rapidly, before flattening out again. We will call it, with some apologies to the English alphabet, the S-shape curve.* The S-shape of this curve is the source of the poverty trap » (p. 10, souligné par l'auteure).

Cette dernière phrase apparaît deux fois, avec une légère variation : « *The sinusoidal character of the curve is at the origin of / generates the poverty trap* » (p. 10 et p. 200). Étonnamment, ces phrases font littéralement de la forme de la courbe la cause du piège, alors que l'on pourrait penser qu'elle en est plutôt la représentation. Nous ne sommes pas loin de la pensée magique. Étroitement liée à cette courbe, l'idée du piège à pauvreté est aussi métaphore qui recourt à un champ sémantique voisin. L'individu peut « plonger » (*plunge*) dans ces pièges et en rester prisonnier (12 occurrences pour *trapped*, dont 6 pour *trapped in poverty*), ou coincé (*stuck*, p. 43). Cette métaphore est filée par celle des échelles (*ladders*) qui permettent d'échapper à ces pièges :

> « *As he* [*Jeffrey Sachs*] *sees it, there are healthbased* poverty traps, *but there are also* ladders *we can give to the poor to help them* escape from these traps. *If the poor cannot afford these* ladders, *the rest of us should help them out* » (p. 46, souligné par l'auteure).

> « *The* ladders to get out of the poverty trap *exist but are* not always in the right place, *and people do not seem to know* how to step onto them *or even want to do so* » (p. 50, souligné par l'auteure).

Il est aussi question de relâcher le piège (*to set the trap loose*, p. 42), de le casser (*break the trap*, p. 234), d'en échapper et d'en sortir (*escape from the trap[s]*, p. 10 et 46 ; *getting out of the trap* : titre, p. 200). Comme l'a montré McCloskey (1983 : 502), « l'économie est lourdement métaphorique » et ses « modèles sont des métaphores ». *Poor Economics* en est une illustration, élevant l'art de la métaphore et de l'anecdote à un niveau inaccoutumé.

Une saisissante profusion d'anecdotes

L'histoire d'Ibu Tina n'est qu'une anecdote parmi tant d'autres. Pourtant, un corps de doctrine largement hostile à l'anecdote se fait jour quand on se penche sur les 10 occurrences du concept d'*anecdote* (ou de l'adjectif *anecdotal*) dans le texte (encadré 1).

La doctrine : les données des RCT, antidotes à des anecdotes fallacieuses

Encadré 1 : Inventaire des occurrences du terme *anecdote* dans *Poor Economics*[9]

1. « *If the poor appear at all, it is usually as the* dramatis personae of some uplifting anecdote *or* tragic episode, *to be* admired *or* pitied » (p. viii).
2. « *We* need evidence. *But unfortunately, the kind of data usually used to answer the big questions does not inspire confidence. There is* never a shortage of compelling anecdotes, *and it is* always possible to find at least one to support any position » (p. 4).
3. « *However, there are now* a number of careful experiments that suggest that such anecdotes are oversold » (p. 57).
4. « *For all the* individual anecdotes *of fruit sellers turning into fruit magnates that can be found on the various Web sites of microfinance institutions*, there are still many poor fruit sellers in Chennai » (p. 159).
5. « *We met a prominent Silicon Valley venture capitalist and investor, and supporter of microcredit […], who told us that he needed* no more evidence. He had seen enough "anecdotal data" to know the truth » (p. 168).
6. « Anecdotal data does not help with the skeptics out there, *including large sections of governments everywhere* […] » (p. 169).
7. « The anecdotes […] did little to help them out. *One reason the MFIs were lacking a* powerful argument *in their defense is that they had been* reluctant to gather rigorous evidence to prove their impact […] » (p. 181).
8. « Anecdotal data is not truth or evidence » (p. 168).
9. « *The MFIs* responded to the evidence from the two [RCT] studies […] with six anecdotes *on successful borrowers* » (p. 172).
10. « *A* study […] *shows that this role* […] goes beyond this particular anecdote » (p. 228).

9. Dans les citations, c'est l'auteure qui souligne.

Dans *Poor Economics*, l'anecdote apparaît ouvertement et fondamentalement comme trompeuse. Elle est présentée comme un repoussoir, manipulable à l'envi (« *it is always possible to find at least one* [*anecdote*] *to support any position* »), permettant de survendre (*oversold*) certaines opinions, suscitant diverses émotions : la fascination (*compelling*), le réconfort (*upflifting*), l'admiration (*admired*) ou la pitié (*pitied*) et non la raison. Aussi est-elle systématiquement opposée au registre de la preuve (*evidence*, *to prove*) et de la vérité (*truth*) issues des RCT (*experiments*, *studies*) qui sont, eux, associés à la rigueur (*rigourous*) et la minutie (*careful*). Au mieux, de manière classique dans la littérature académique, l'anecdote relève du cas individuel (*individual*), particulier (*particular*) devant être dépassé (*goes beyond*) pour accéder à la généralité (non pas les quelques individus devenus magnats de la vente de fruits, mais la majorité des vendeurs de fruits). Mais certains acteurs et les institutions de microfinance résistent à la preuve scientifique (*relectant to gather rigourous evidence*). Ils abusent des anecdotes qu'ils prennent fallacieusement pour preuve. Mais la morale de l'histoire, contée p. 168, est heureusement la suivante : pour qui n'écoute pas les sirènes des « données anecdotiques », l'anecdote est de peu de secours (*did little to help them out*), nourrit un scepticisme (*skeptics*) fondé et ne permet pas en dernière instance d'emporter la conviction (*lacked a powerful argument*). Seules les RCT le permettent : sont ainsi évoquées les « *demonstrations of the persuasive power of a successful randomized experiment* » (p. 78). Les « données anecdotiques » s'avèrent un leurre.

La pratique des anecdotes : un foisonnement paradoxal

McCloskey (1983 : 482) fait état d'un phénomène qui semble s'appliquer aux anecdotes dans *Poor Economics* :

> « *ECONOMISTS DO NOT FOLLOW* [*sic*] *the laws of enquiry their methodologies lay down.* […] *Their genuine, workaday rhetoric, the way they argue inside their heads or their seminar rooms, diverges from the official rhetoric.* »

Car, si les chiffres sont ubiquitaires dans le livre, ils ne sont pas mobilisés comme antidotes à des anecdotes. Ils fonctionnent, au contraire, en symbiose avec les anecdotes rapportées par les auteurs. Si l'on examine la structure de l'ouvrage, il apparaît que de nombreux chapitres s'ouvrent sur une anecdote, le plus souvent nominative. Le titre du chap. 5 est même associé au nom du protagoniste de l'une d'entre elles : « *Pak Sudarno's Big Family* ». Elles ne sont pas qu'une entrée en matière, on les retrouve au fil des développements de chaque chapitre. Leur présence systématique n'est pas anecdotique. Les anecdotes les plus fréquentes et les plus saisissantes concernent les pauvres (encadré 2). Ce sont bien les « *dramatis personae* » du livre.

Encadré 2 : Douze anecdotes nominatives mettant en scène des pauvres dans *Poor Economics*

1. L'histoire de Pak Solhin (Indonésie) : ouvrier agricole au chômage, pêcheur occasionnel, thème du piège à pauvreté nutritionnelle, 16 occurrences.
2. L'histoire d'Allal Ben Sedan (Maroc) : petit éleveur qui ne voulait pas du microcrédit, 13 occurrences.
3. L'histoire de Xu Aihua (Chine) : personne pauvre devenue entrepreneuse à succès, 12 occurrences.
4. L'histoire de Pak Sudarno (Indonésie) : chiffonnier, père de neuf enfants, thème de la démographie, 10 occurrences.
5. L'histoire de Michael et Anna Modimba (Kenya) : producteurs de maïs, thème de l'accès aux engrais, 10 occurrences.
6. L'histoire de Kennedy (Kenya) : agriculteur, thème de l'utilisation des engrais, 9 occurrences.
7. L'histoire de Jennifer Auma (Kenya) : vend des céréales et des légumineuses au marché, thème de l'épargne des pauvres et des tontines, 8 occurrences.
8. L'histoire d'Ibu Emptat (Indonésie) : épouse d'un vannier, thème du piège à pauvreté sanitaire, 7 occurrences.
9. L'histoire de Shantarama (Inde) : veuve et mère de six enfants, thème de l'absentéisme scolaire, 5 occurrences.
10. L'histoire de Wycliffe Otieno (Kenya) : agriculteur, thème de l'accès aux engrais, 5 occurrences.
11. L'histoire de Pak Awan (Indonésie) : ouvrier du bâtiment au chômage, ouverture d'une échoppe faute de mieux, thème de l'entrepreneuriat forcé, 4 occurrences.
12. L'histoire d'Oucha Mbarbk (Maroc) : ouvrier agricole et du bâtiment occasionnel, se privant de nourriture pour acheter un téléviseur, thème de la préférence pour les loisirs, 2 occurrences.

Des anecdotes obéissant aux principes du marketing social et du *storytelling*

Ces pauvres incarnent chacun un problème emblématique de l'économie de la pauvreté. Par exemple, Pak Sohlin incarne le piège à pauvreté lié à une nourriture insuffisante : « *Pak Solhin* […] *once explained to us exactly how such a poverty trap worked* » (p. 23).

Son histoire est détaillée en 577 mots, soit près de trois pages. Certains de ces pauvres sont des personnages récurrents : leur histoire est relatée dans l'un des

chapitres et ils reviennent dans d'autres épisodes (16 apparitions pour Pak Sohlin). Ces histoires répondent aux principes du *storytelling* : elles sont *personnelles, vraies, simples, informatives, concrètes, circonstanciées, émotionnelles et traduisibles en actes* (*actionable*) (FEW, 2009). Comme le disent Banerjee et Duflo : « *The point is simple: Talking about the problems of the world without talking about some accessible solutions is the way to paralysis rather than progress* » (p. 5). Ce *storytelling* vise à persuader de l'intérêt des expérimentations et des solutions qu'elles valident, de faire appel aux financements de divers organismes comme aux dons. Il participe du marketing social (LEE *et al.*, 2011 : 1).

Anecdotes et efficacité caritative : un effet démontré expérimentalement

De par les programmes qu'ils évaluent, Banerjee et Duflo connaissent très bien les techniques de marketing social. Ces techniques sont en effet appliquées à l'aide au développement et aux politiques de santé dans les pays du Sud, en particulier en Inde. Le paternalisme *nudge* revendiqué par les auteurs en est d'ailleurs une composante. L'impact de ce marketing sur les dons philanthropiques (*charitable giving*) est un domaine d'étude privilégié des expérimentations randomisées contrôlées aux États-Unis (LIST et LUCKING-REILEY, 2002).

Banerjee et Duflo mentionnent dès la deuxième page de l'ouvrage – indice de l'importance du sujet – les résultats d'une expérimentation portant sur l'efficacité de deux types d'appels au don pour une organisation caritative. Dans le premier groupe, l'appel au don repose sur des données générales et abstraites sur la pénurie alimentaire générée par une moindre pluviométrie et emploie un vocabulaire scientifique (du pur *logos*). Dans le second, qui va récolter plus de deux fois plus de fonds, l'appel est personnalisé et émotionnel (*pathos*), il met en scène la petite Rokya, 7 ans, « désespérément pauvre et menacée par la faim », dont il est possible, grâce aux dons, de changer la vie. L'usage d'anecdotes personnalisées par Banerjee et Duflo y trouve l'une de ses raisons d'être. Il en est d'autres.

Des anecdotes témoignant de l'*ethos* des auteurs : proximité, familiarité, crédibilité

Les anecdotes de *Poor Economics* sont personnelles à double titre. Elles sont personnalisées par le nom et le prénom du personnage principal. Elles sont personnelles, car il s'agit, presque toujours, de personnes que Banerjee et Duflo ont rencontrées sur le terrain. Aussi, ces anecdotes témoignent-elles de la proximité des auteurs à l'égard des pauvres, une composante essentielle de leur *ethos* dans le livre. Dès la préface, ils se présentent comme de modestes invités des pauvres auxquels ils sont reconnaissants (p. viii). Le témoignage de leur passage parmi les pauvres est un *leitmotiv* : on compte ainsi 34 occurrences

de « [*name*] *we* [*the authors*] *met* ». Ce passage sur le terrain est mis en avant, là encore, dans la préface :

> « *We are academics, and like most academics we formulate theories and stare at data. But the nature of the work we do has meant that we have* also spent months, spread over many years, on the ground *working with NGO activists and government bureaucrats, health workers and microlenders. This has taken us to* the back alleys and villages where the poor live, asking questions, looking for data » (p. vii, souligné par l'auteure).

Voilà qui confère aux auteurs un « avantage rhétorique comparatif » face à des rivaux comme Acemoglu et Robinson (voir section « Deux schémas rhétoriques aux puissants effets épistémiques et persuasifs »). Peu d'économistes *mainstream* peuvent se targuer d'autant de « terrains » que les auteurs de *Poor Economics*. La profession fonctionne d'abord en chambre, par le traitement de données à distance. En revanche, dans une conception plus exigeante du terrain (celle des autres sciences sociales), ces terrains apparaissent comme trop rapides et peu rigoureux (Jatteau, 2013 et section « Remarques finales : des récits persuasifs, mais pauvres »).

Par ailleurs, certaines anecdotes concernent aussi les auteurs eux-mêmes. Les deux auteurs sont identifiés par leur seul prénom, Esther (47 occurrences) et Abhijit (38 occurrences). C'est le cas dès les premières lignes du livre :

> « *Esther was six when she read in a comic book on Mother Theresa that the city then called Calcutta was so crowded that each person had only 10 square feet to live in. She had a vision of a vast checkerboard of a city, with 10-foot squares marked out on the ground, each with a human pawn, as it were, huddled into it. She wondered what she could do about it* […] »
>
> « *At six, Abhijit knew where the poor lived. They lived in little ramshackle houses behind his home in Calcutta. Their children always seemed to have lots of time to play, and they could beat him at any sport: When he went down to play marbles with them, the marbles would always end up in the pockets of their ragged shorts. He was jealous* » (p. vii).

Un tel procédé relève de la *captatio benevolentiae* (Dokova, 2016), un processus oratoire classique visant à attirer la bienveillance et la sympathie d'un auditoire dès l'exorde (début du discours). Il crée une « familiarité avec l'auditoire ». Cette familiarité est renforcée ici par le fait que les auteurs sont présentés comme de jeunes enfants naïfs (gage de modestie), en proie à des passions très humaines comme la jalousie. Ce procédé est récurrent :

> « *When a puzzled six-year-old Abhijit asked his Bengali aunt* […] » (p. 70).
>
> « *Abhijit was falling behind in his schoolwork in first grade* […] » (p. 90).
>
> « *The school Abhijit went to in Calcutta* […] » (p. 94).

Cette familiarité est entretenue par la référence à d'autres membres de la famille d'Abhijit. Outre sa tante, ce sont son père, sa mère et son grand-père qui font eux aussi l'objet d'anecdotes (p. 70 et 183).

L'*ethos* des auteurs est complété par leur présentation académique (*ethos* pré-discursif[10]). La familiarité induite ci-dessus est assortie à de multiples gages de crédibilité et d'autorité scientifique. Cette présentation détaille le *cursus honorum* des auteurs et leur capital académique et symbolique exceptionnel (les diplômes et les prix les plus prestigieux). Dans le reste du livre, Esther et Abhijit sont également représentés comme des adultes, des scientifiques expérimentés et expérimentateurs :

> « *But several simultaneous experiments that Esther, Pascaline Dupas, and Michael Kremer conducted* […] » (p. 114).
>
> « *To evaluate it, Esther compared* […] » (p. 81).
>
> « *In a study with Udry, Esther found* […] » (p. 125).
>
> « *Abhijit with two Chinese-born co-authors, Nancy Qian* […] *and Xin Meng* […] » (p. 120).

On remarquera que les co-auteurs sont désignés par leur nom de famille, contrairement à « Abhijit » et « Esther ». Mises bout à bout, ces anecdotes personnelles portant sur les pauvres, comme sur les auteurs et leurs proches, dessinent donc un remarquable portrait de l'*ethos* des auteurs. Ils sont bienveillants à l'égard des pauvres, familiers de leur situation, ils sont modestes et accessibles alors même qu'ils ont accompli un parcours d'excellence, parcours qui garantit la haute autorité de leurs propos.

Anecdotes, didacticité et distinction

Les anecdotes font depuis longtemps partie intégrante de l'arsenal pédagogique (Stock, 1993 ; Ford, 2002). Figurant parmi les « procédés d'exemplification », elles remplissent une fonction de didacticité : « une intention réelle ou simulée […] d'apporter à l'autre des savoirs nouveaux » (Beacco et Moirand, 1995 : 33). Ces petites histoires donnent de la chair aux données, permettent d'attirer et de retenir l'attention du lecteur tout au long d'un ouvrage relativement long. Les membres du J-PAL, parce qu'ils s'adressent régulièrement à des publics non académiques (responsables d'organisation non gouvernementale [ONG], d'organisations internationales, décideurs politiques, etc.), ont développé des compétences didactiques étendues. Kuhn (1962 : 187) a montré la centralité des *exempla* dans l'apprentissage d'un paradigme et la différenciation des communautés scientifiques : « *More than other sorts of components of the disciplinary matrix, differences between sets of exemplars provide the community fine-structure of science* ». Dans *La Rhétorique* (livre II, chap. 20), Aristote

10. Il est annexé au discours : formellement, ce ne sont pas les auteurs qui se présentent. « L'*ethos* pré-discursif renvoie quant à lui, à la réputation de l'orateur […] [Il] est à l'œuvre dans un système fondé sur l'*auctoritas* » (Duteil-Mougel, 2005 : 4).

distingue les groupes d'exemples relatant des faits historiques et des exemples fabriqués. Ici les exemples sont des histoires vraies, et non des fables (du troc et autres robinsonnades) ou des exemples imaginaires, un mode de narration très fréquent dans les manuels d'économie (JALLAIS, 2018).

Les anecdotes authentiques de *Poor Economics*, tout particulièrement celles qui touchent aux pauvres, distinguent donc la communauté épistémique du J-PAL au sein du champ disciplinaire. Elles les différencient aussi dans le champ des économistes « praticiens du développement ». Comparons *Poor Economics* à d'autres ouvrages de vulgarisation, écrits par des économistes qui, en raison de leurs fonctions dans les organisations internationales, ont réalisé de multiples déplacements sur le terrain. Chez STIGLITZ (2006), on compte très peu d'anecdotes. SACHS (2005) fait, lui, largement appel aux anecdotes personnelles dans *The End of Poverty.* Mais celles-ci portent essentiellement sur des *president* (70 occurrences), *minister* (50), *leader* (106), *secretary-general* (17) et autres *director* (14). C'est très symptomatique de sa vision « *top-down* » de conseiller du prince. Les pauvres rencontrés dans les *Millenium Villages* sont plus marginalement les sujets – anonymes – de l'histoire. EASTERLY (2001), dans *The Elusive Quest for Growth*, mentionne quelques anecdotes sur les pauvres, notamment dans l'« *Intermezzo : Leila's Story* ». Certains de ces récits sont tirés de ses déplacements sur le terrain, d'autres des médias. Fait remarquable, les récits d'Easterly concernent au premier chef des pays entiers (Ghana, Côte d'Ivoire, Inde, Chine, etc.). Ces grands récits structurent le livre, un point de convergence avec Sachs (narrations sur la Bolivie, la Pologne, la Russie, la Chine, etc.) et de divergence avec Banerjee et Duflo. Eux se concentrent exclusivement sur de « petites » histoires personnelles. Contrairement à ces deux derniers, aucun de ces économistes n'élabore de doctrine hostile aux anecdotes : il n'y a aucune occurrence de cette notion dans Sachs, une dans Stiglitz, cinq dans Easterly, mais sans connotation négative.

La discrète fonction heuristique des anecdotes

Chez Banerjee et Duflo, le rôle des anecdotes ne se cantonne pas à ces dimensions éthiques, rhétoriques et pédagogiques. En pratique, les anecdotes jouent un rôle heuristique feutré, mais crucial. Dans une interview accordée à France Culture (6 janvier 2012), Esther Duflo expliquait que ces histoires « permettent de mieux comprendre les données statistiques ». Dans *Poor Economics*, Banerjee et Duflo rapportent que ces histoires leur permettent de « tisser une histoire cohérente » :

> « *Many stories were shared with us. Back in our offices,* remembering these stories and analyzing the data, *we were both fascinated and confused, struggling to fit what we were hearing and seeing into the simple models that* […] *professional development economists and policy makers use to think about the lives of the poor.* […] *This book comes out of that interchange ; it represents our attempt to* knit together a coherent story *of how really poor people live their lives* » (p. viii, souligné par l'auteure).

Le recours au récit pour comprendre les données illustre l'incomplétude foncière du chiffre issu des expérimentations aléatoires dès lors qu'il s'agit de comprendre le pourquoi de résultats surprenants et donner un sens aux données. En effet, comme je l'ai démontré ailleurs (LABROUSSE, 2017) : « Il est malaisé de comprendre le chemin de causalité (comment ? par quels mécanismes ?) conduisant à un ensemble particulier de résultats observés. En effet, hormis les cas de mono-causalité simple (une cause entraîne un effet, sans effet rétroactif sur la cause), les expérimentations randomisées fournissent des preuves d'efficacité (un effet est observé) plutôt que de causalité (quels mécanismes ont généré cet effet ?)[11] […] En présence d'une causalité complexe, cumulative, multifactorielle et non linéaire, les chaînes causales deviennent une sorte de boîte noire pour les expérimentateurs. »

Toutefois, la mise en récit dans *Poor Economics* est trop mince et trop imprégnée d'hypothèses *a prior*i pour mettre au jour des causalités pertinentes. Elle est souvent trompeuse. Par exemple, ces discours narratifs légers soulignent les défaillances des pauvres, comme leur tendance à dépenser plutôt qu'à économiser et à dépenser pour de « mauvaises » choses, comme les rituels cérémoniels ou les achats de thé (p. 37, 171, 183-204). Des travaux ethnographiques sur l'Inde montrent que les échoppes à thé constituent des lieux stratégiques en termes de réseaux et de collecte d'informations, que les cadeaux cérémoniels représentent une forme d'épargne relationnelle : « Les ménages, y compris ceux au bas de la pyramide, épargnent, en ce sens qu'ils stockent, accumulent et font circuler de la valeur. Mais cela se fait par des formes particulières de médiation [comme les cérémonies] qui permettent aux épargnants de forger ou d'entretenir des relations sociales et émotionnelles, de garder le contrôle sur la valeur […] Les gens préfèrent créer de la valeur – et épargner – en investissant dans leurs réseaux sociaux plutôt que de bloquer leurs avoirs sur un compte bancaire. […] Affirmer que les pauvres n'ont pas la notion du temps, manquent de maîtrise de soi, ou de discipline pour résister à la pression sociale […] montre une incompréhension totale des dynamiques sociales et économiques locales » (GUÉRIN *et al.*, 2019 : 1 et 13). Ce récit alternatif, bien plus solide, repose notamment sur l'observation dense des cérémonies et l'examen scrupuleux des carnets dans lesquels les pauvres tiennent une comptabilité détaillée des dons et contre-dons pour chaque cérémonie, ainsi que sur des entretiens approfondis revenant sur chaque transaction. Dans la RCT *Proempleo*, des entretiens qualitatifs de suivi ont permis d'éviter une interprétation complètement erronée des résultats expérimentaux (Ravallion, chap. 1, ce volume). Dans le cas de la RCT *Al Amana* le matériel ethnographique a remis en question le discours du J-PAL sur la demande en microcrédit (MORVANT-ROUX *et al.*, 2014). Ces « menus récits » des randomistes s'opposent à la « description dense » des ethnographes (GEERTZ, 1973).

11. Sur la distinction entre la preuve de ce qui fait la différence et la preuve d'un mécanisme (c'est-à-dire la causalité), voir BERRIET-SOLLIEC *et al.* (2014).

Deux schémas rhétoriques aux puissants effets épistémiques et persuasifs

Loin de cette critique socio-économique, la rhétorique de *Poor Economics* est confortablement ancrée dans l'économie *mainstream*, le seul type d'économie référencé. Cette rhétorique génère un « avantage persuasif » dans la concurrence entre les preuves au sein des courants dominants. Elle permet au livre de combattre les explications procédant d'écoles de pensée concurrentes. Deux schémas rhétoriques transversaux sont particulièrement efficaces ici. D'abord, la « rhétorique de la voie médiane entre deux extrêmes » renforce la posture raisonnable et a-idéologique du J-PAL. Elle permet ainsi de décrédibiliser des concurrents de la Ivy League^{+} comme Sachs et Easterly. Elle polarise le débat et occulte le large éventail des approches restantes. Ensuite, la « rhétorique des petites mesures aux grands effets » permet de majorer le micro et de minorer le macro : elle légitime la posture du J-PAL et disqualifie l'économie politique et l'institutionnalisme d'ACEMOGLU et ROBINSON (2006).

La voie médiane entre deux extrêmes : bon sens, objectivité et cadrage manipulatoire

Deux économistes influents incarnent les deux pôles opposés de l'économie du développement : Jeffrey Sachs (39 occurrences) d'un côté et William Easterly (33 occurrences) de l'autre. Ce sont des antithèses, figures d'oppositions symétriques dont Aristote (*La Rhétorique*, livre III, chap. XIX) dit : « Ce genre de style est agréable, parce que les contraires sont très reconnaissables et que les idées mises en parallèle n'en sont que plus faciles à saisir. » En l'occurrence, Sachs et Easterly sont décrits comme ayant chacun une réponse universelle à tout, mais apparaissent opposés sur tout. Leur opposition est dépeinte dès l'introduction comme une guerre picrocholine entre deux quartiers de Manhattan :

> « Jeffrey Sachs, *adviser to the United Nations, director of the Earth Institute at* Columbia University in New York City […], *has an answer to all these questions* […].
>
> *But then there are others, equally vocal, who believe that all of Sachs's answers are wrong. William Easterly, who battles Sachs from New York University at the other end of Manhattan, has become one of the most influential anti-aid public figures* […] » (p. 2, souligné par l'auteure).

Pour les auteurs, ces désaccords sont d'ordre idéologique :

> *« It is no accident that Sachs and Easterly have radically opposite views on whether bed nets should be sold or given away. The positions that most rich country experts take on issues related to development aid or poverty tend to be colored by their specific worldviews* […] *on the left of the political spectrum, Jeff Sachs (along with the UN, the*

> *World Health Organization, and a good part of the aid establishment) wants to spend more on aid* […]. *On the right, Easterly, along with Moyo, the American Enterprise Institute, and many others, oppose aid* […] » (p. 8-9).

Cette coloration idéologique discrédite les deux contradicteurs. Par contraste, Banerjee et Duflo vont apparaître comme la voix de la raison et du bon sens. Ils ne sont pas dans l'idéologie et la position de principe, mais dans l'objectivité et la preuve empirique. Ils abordent les problèmes concrètement, un à un (encadré 3).

Encadré 3 : Loin des réponses toutes faites à la Sachs et Easterly, des réponses concrètes

« *This book is an invitation to think again, again : to turn away from the feeling that the fight against poverty is* too overwhelming, *and to start to think of* the challenge as a set of concrete problems *that* […] *can be* solved one at a time » (p. 1-2, souligné par l'auteure).

« *There are in fact* answers—*indeed, this whole book is in the form of an* extended answer—*it is just that they are* not the kind of sweeping answers that Sachs and Easterly favor » (p. 3, souligné par l'auteure).

« *This is why it is really helpful to think in terms of* concrete problems, *which can have* specific answers » (p. 5, souligné par l'auteure).

« *These questions can be* answered, *but the answers are by no means obvious. Yet* many "experts" take strong positions on them that have little to do with evidence » (p. 6, souligné par l'auteure).

« *This radical shift in perspective*, away from the universal answers, *required us to* step out of the office and look more carefully at the world » (p. 13, souligné par l'auteure).

« *There is* no general rule here […]. *It is the body of knowledge that grows out of* each specific answer *and the understanding that goes into* those answers *that give us the* best shot *at, one day,* ending poverty » (p. 14, souligné par l'auteure).

« […] *although we have* no magic bullets *to eradicate poverty*, no one-shot cure-all, *we do* know *a number of things about how* to improve the lives of the poor » (p. 267, souligné par l'auteure).

Les auteurs apparaissent ainsi plus méticuleux, plus réalistes et plus modestes : ils disposent de solutions, non des solutions miracle, universelles[12], mais des solutions adaptées à chaque problème. Ils sont dans le concret et non la spéculation. Ce schème rhétorique complète le portrait éthique des

12. On peut néanmoins s'interroger ce point : « même si Duflo refuse d'évaluer l'utilité (ou la nocivité) de l'"aide" ou de l'"éducation" en général, le terrain de ce qui est "bon" ou du "mauvais" en soi se déplace simplement à un niveau d'analyse plus concret : pour le J-PAL, certains dispositifs micro-sociaux peuvent être intrinsèquement bons ou mauvais pour tous les pauvres dans chaque domaine d'étude (éducation, nutrition, etc.) » (Labrousse, 2016 : 286).

auteurs. Il correspond à un énoncé de cadrage manipulatoire, par fausse alternative entre deux extrêmes idéologiques. Les auteurs laissent ainsi dans l'ombre toutes les approches ne relevant pas de ces deux formes de néolibéralisme[13]. C'est le cas des approches de Rodrik ou Stiglitz (jamais cités), des nombreux courants d'économie politique qui ne s'inscrivent pas dans le *mainstream*, comme de l'économie du développement classique (Myrdal, Hirschman, Boserup etc.).

Petites causes, grands effets : des oxymores en défense du « tout micro »

Dans l'extrait ci-dessous, le schème précédent se combine à un second schème structurant, une rhétorique du « *think small to solve global problems* » :

> « *We are often asked why we do what we do: "Why bother?"* These are the "small" questions. William Easterly, for one, criticized randomized control trials *(RCT) on his blog in these terms: "RCTs are infeasible for many of the* big questions *in development, like the* economy-wide *effects of good institutions or good* macroeconomic *policies." Then, he concluded that "embracing RCTs has led development researchers to* lower *their ambitions." This statement was a good reflection of an institutionalist view that has strong currency in development economics today.* […] *It follows (or so it is assumed) that* "big questions" require "big answers"—social revolutions, *such as a transition to effective democracy.* At the other extreme, Jeffrey Sachs *sees corruption, perhaps not surprisingly, as a poverty trap: Poverty causes corruption, and corruption causes poverty* » (p. 236, souligné par l'auteure).

Pour contrer la rhétorique des « grandes questions – grandes réponses », les auteurs vont déployer la rhétorique des « petites causes – grands effets ». Ce schème est récurrent. Il apparaît dès la préface :

> « *the* small costs, *the* small *barriers, and the* small *mistakes that most of us do not think twice about* loom large in the lives of those who have very little.
>
> *It is not easy to escape from poverty, but a sense of possibility and a* little bit of well-targeted help *(a* piece *of information, a* little nudge*) can sometimes have* surprisingly large effects.
>
> *On the other hand, misplaced expectations, the lack of faith where it is needed, and* seemingly minor hurdles can be devastating. *A* push on the right lever *can make a* huge difference, *but it is often difficult to know where that lever is. Above all, it is clear that* no single lever will solve every problem » (préface, p. x, souligné par l'auteure).

13. Néolibéralisme centralisateur pour Sachs et décentralisateur pour Easterly (tradition autrichienne).

Il se retrouve dans les huit occurrences de *small changes*, dont :

> « small changes, *we believe, can sometimes end in a* quiet revolution » (p. 237, souligné par l'auteure).
>
> « *What is not recognized as often, however, is* how important the effect *of* seemingly very small changes *can be* » (p. 246, souligné par l'auteure).
>
> « *A small change in the rules changed everything* » (p. 249).
>
> « *Don't let the* apparent modesty *of the enterprise fool you:* Small changes can have big effects » (p. 272, souligné par l'auteure).

Il est répété avec l'adjectif *minor* :

> « Seemingly minor *interventions can make a* significant difference » (p. 253, souligné par l'auteure).
>
> « Seemingly minor *hurdles can be* devastating » (préface, p. x, souligné par l'auteure).
>
> « *A* seemingly minor *technical fix,* involving no major political battle, changed the way *in which the voice of the poor was taken into account* » (p. 247, souligné par l'auteure).

On en trouve une variation avec l'adjectif *incremental* :

> « *We are not "lowering our ambitions": Incremental progress and the accumulation of these small changes, we believe, can sometimes end in a quiet revolution* » (p. 237).
>
> « *These changes will be incremental, but they will sustain and build on themselves. They can be the start of a quiet revolution* » (p. 265).

La même idée est exprimée avec *step* : *small steps* (1 occurrence) et *step-by-step* (1 occurrence), *first step* (9 occurrences), *stepping stone* (2 occurrences), qui sont opposés à *extreme steps* (1 occurrence), *drastic steps* (1 occurrence). Elle se retrouve dans la métaphore des briques qui, une à une, vont permettre de construire une maison. *Brick by brick* est même le titre du chap. 8. On dénombre cinq occurrences de cette expression et quatorze de *brick*.

Ce *leitmotiv* fait appel à plusieurs figures de style. Des *métaphores évocatrices*, comme celle des petits pas, marches et autres marche-pied dans l'avancée vers le progrès ou des briques construisant la maison. Surtout, Banerjee et Duflo usent systématiquement d'oxymores : *quiet revolution*, *small/big*, *minor/significant* révélant un paradoxe apparent (« *seemingly* »). L'oxymore permet de décrire une situation de manière inattendue, a priori inconcevable. Elle s'inscrit dans une stratégie de la surprise et du pathos (Monte, 2007). La contradiction dans les termes de ces oxymores n'est qu'apparente. Grâce à la co-occurrence presque systématique des modalisations *seemingly* et *apparent*, les auteurs inhibent le jugement de contradiction pour mieux persuader.

Une autre propriété de l'oxymore est de prendre le contrepied d'une *doxa* (Monte, 2007). C'est la *doxa* des approches macroéconomiques et institutionnalistes qui doit être écartée ici.

Ces oxymores sont des *oxymores de combat* : la microéconomie empirique part à l'assaut de la « forteresse macroéconomique théorique », selon Angrist et Pischke (2010), économistes proches de Banerjee et Duflo. La métaphore des briques est également au service de cette idée. Duflo avait déjà justifié le primat du micro sur le macro en employant la métaphore du Meccano, un jeu de construction :

> « *Using macroeconomic data* […] *leads to a stalemate.* […] *the macroeconomic model is constructed like a Meccano set, based on microeconomic building blocks* […] *In any case, the basic elements are microeconomic elements* » (Duflo, 2009 : 73-74).

Le macro ne serait que la somme des comportements micro, comme la maison ne serait que la somme de ses briques. Ce réductionnisme est typique d'un *sophisme de composition* (Labrousse, 2010 et 2016). L'économie politique institutionnelle d'Acemoglu et de Robinson est également visée :

> « *Our* colleague Daron Acemoglu, *and his long-term coauthor,* Harvard's James Robinson, *are two of the most thoughtful exponents of the rather* melancholy view, *active in economics today, that until political institutions are fixed, countries cannot really develop, but institutions are hard to fix. Both* political scientists and economists *typically think of institutions at a very high level. They have in mind, if you like,* institutions in capital letters—*economic INSTITUTIONS like property rights, or tax systems ; political INSTITUTIONS like democracy or autocracy, centralized or decentralized power, universal or limited suffrage* » (p. 236-237, souligné par l'auteure).

L'usage répété de ce contraste frappant entre des concepts clé en capitales (le macro, l'abstrait, le grand) et en bas-de-casse (le micro, le concret, le spécifique) vient visualiser et renforcer le procédé :

> « *To really understand the effect of institutions on the lives of the poor, what is needed is a shift in perspective from INSTITUTIONS in capital letters to institutions in lower case—the "view from below"* » (p. 243).

> « *The focus on the broad INSTITUTIONS as a necessary and sufficient condition for anything good to happen is somewhat misplaced. The political constraints are real, and they make it difficult to find big solutions to big problems. But there is considerable slack to improve institutions and policy at the margin* » (p. 263, souligné par l'auteure).

On compte cinq occurrences de ce contraste typographique. Il est également décliné avec « aide » plutôt qu'« Aide ». Duflo et Banerjee partent ainsi en

croisade « *AGAINST POLITICAL ECONOMY* » (titre de niveau 4), dont ils donnent une curieuse définition[14] :

> « *Political economy is the view (embraced, as we have seen, by a number of development scholars) that politics has primacy over economics: Institutions define and limit the scope of economic policy* » (p. 252).

Ils moquaient l'opposition entre les deux extrémités de Manhattan de Sachs et d'Easterly. Ils sont ici les protagonistes d'une lutte au sein du département d'économie du MIT (avec Acemoglu) et avec le département du gouvernement de l'université voisine de Harvard (Robinson), les deux n'étant séparés que de quelques kilomètres à Cambridge (Massachusetts). L'objectif des auteurs est de persuader les lecteurs que leur approche consistant à « *think small* pour lutter contre la pauvreté mondiale » est la meilleure qui soit. Ces enjeux de persuasion sont mentionnés au début de l'introduction et dans sa dernière phrase, ce qui en signale l'importance :

> « *The problem* [*of world poverty*] *seems* too big, too intractable. Our goal with this book *is to persuade you not to* » (p. 1, souligné par l'auteure).
>
> « *We hope to persuade you that our* patient, step-by-step *approach is not only a* more effective way to fight poverty […] » (p. 15, souligné par l'auteure).

Ce schème légitime la multiplication d'expérimentations par le J-PAL. Le développement est réduit à l'implémentation d'une série de petits dispositifs visant à infléchir les conduites au niveau individuel et groupal. Des coups de pouce *nudge* donnent l'impulsion qui permet de modifier les incitations (30 occurrences d'*incentive**) et d'aiguiller vers les bonnes conduites. Cette rhétorique du « *think small* » permet d'éluder des questions brûlantes : augmentation des inégalités, déséquilibre des rapports de force internationaux, etc. Ce qui est hors champ est révélateur. On compte 5 occurrences cumulées seulement pour *structure** ou *structural*. Trois portent sur des microstructures (*the structure of the program, life structured by goals, the structure of banks*). Les deux autres dénient l'importance des macro-structures[15]. Les éléments qui restent en dehors du cadre du livre sont révélateurs. Il n'y a que cinq occurrences cumulatives de *structure* ou *structural*. Trois de ces occurrences concernent des micro-structures (*the structure of the program, life structured by goals, the structure of banks*). Les deux autres occurrences nient l'importance des macrostructures :

14. Pour plus de détails, voir Labrousse (2016).

15. Ce mépris des réformes politiques de grande envergure pourrait expliquer la popularité de *Poor Economics* dans les cercles philanthropiques : il légitime le travail des (méga)fondations, sans pour autant s'attaquer aux problèmes des inégalités, de l'évasion fiscale (allant de pair avec la réduction des budgets sociaux) et de la puissance extractive de nombreuses multinationales dans les chaînes globales de valeur, toutes choses qui pourraient embarrasser certains milliardaires. Voir par exemple Kohl-Arenas (2016 : 16-17), sur la manière dont la philanthropie a séparé « les questions de production, de travail, et l'institution d'inégalités structurelles des explications morales et comportementales de la pauvreté ».

> « *What these two examples (the nurses and the school committees) illustrate is that* large-scale waste and policy failure *often happen* not because of any deep structural problem » (p. 261, souligné par l'auteure).
>
> « *It is possible to improve governance and policy without changing the existing social and political structures* » (p. 270).

Nous observons une configuration comparable pour le *macro* (quatre occurrences seulement : p. 165 et 172, et deux occurrences mettant en avant l'expérimentation fondatrice du J-PAL sur les vermifuges, l'une des rares expérimentations à avoir été appliquée à des populations importantes) :

> « *We may not have much to say about* macroeconomic policies *or institutional reform, but don't let the apparent modesty of the enterprise fool you: Small changes can have big effects. Intestinal worms might be the last subject you want to bring up on a hot date, but kids in Kenya who were treated for their worms at school for two years, rather than one* […]*, earned* 20 percent *more as adults every year* […]*. The effect might be lower if deworming became universal: The children lucky enough to have been dewormed may have been in part taking the jobs of others. But to scale this number, note that Kenya's highest sustained per capita growth rate in modern memory was about* 4.5 percent *in 2006–2008. If we could press a* macroeconomic policy lever *that could make that kind of unprecedented growth happen again,* it would still take four years to raise average incomes by the same 20 percent*. And, as it turns out,* no one has such levers » (p. 272, souligné par l'auteure).

Cet argument compare une hausse des revenus, pour une fraction générationnelle de la population, à la croissance de l'ensemble du pays. Ce sophisme de composition est-il plus honnête que les « anecdotes des vendeurs de fruits devenant des magnats du fruit » ? Les auteurs nient l'existence d'importants leviers macroéconomiques. C'est un argument d'autorité dont on peut se demander comment il peut rendre compte par exemple du développement économique de nombreux pays d'Asie – dont la Chine –, ou de l'effet multiplicateur négatif des politiques d'austérité (Christiano *et al.*, 2011) ? Les PAS ont eu un impact majeur sur la vie des personnes démunies, sur les budgets de l'éducation, de la santé, des infrastructures de transport ou sur l'accès à la nourriture. Détailler la vie des pauvres sans évoquer ces sujets impressionne. Nous avons vu (partie « Qu'est-ce que le monde de Kennedy ? Représenter et réduire le champ des possibilités à deux diagrammes ») l'étendue des problèmes laissés hors-champ par les RCT. Ainsi, l'analyse textuelle contribue-t-elle à mettre au jour leur portée limitée.

Ces procédés rhétoriques amplifient le micro et minimisent le macro. Ils sont manipulatoires implicitement sur une fausse alternative, par exclusive, du tout-micro ou tout-macro. On est dans la *concurrence des preuves et non dans la mise en relation des preuves*. Comme le montre Revel (1996 : 12), « le problème n'est pas tant d'opposer un haut et un bas, les grands et les petits, que de reconnaître

qu'une réalité sociale n'est pas la même selon le niveau d'analyse où l'on choisit de se situer ». C'est alors la mise en regard des niveaux d'observation qui est éclairante : micro, méso et macro sont tout aussi essentiels.

Remarques finales : des récits persuasifs, mais pauvres

> « Il y a trois choses qui donnent de la confiance dans l'orateur […] [:] le bon sens, la probité et la bienveillance. »
>
> Aristote (*La Rhétorique*, livre II, chap. 1, paragraphe 5)

Une combinaison originale des trois piliers de la rhétorique

Poor Economics s'appuie sur les trois piliers de la rhétorique dont il propose une combinaison originale. Le *logos*, tout d'abord, avec un discours qui fait appel à la preuve chiffrée, à une argumentation qui réunit tous les attributs de la rationalité scientifique (argumentation démonstrative, résultats expérimentaux, graphiques, appareil bibliographique). Mais ce *logos* est inextricablement entremêlé de *pathos*. Les chiffres sont associés à des anecdotes émotionnelles et certains chiffres comme le 99 cents ont un caractère iconique. Quant aux graphiques, abstraits à première vue, ils sont eux-aussi personnifiés et mis en récit par des anecdotes et narrations graphiques. De plus, les auteurs usent de multiples figures de style (métaphores, synecdoques, métonymies, hypostases, oxymores, etc.) faisant appel au registre des émotions. L'*ethos* n'est pas en reste. Les auteurs apparaissent doués de sagesse (*phronesis*), de vertu et d'excellence (*arete*), de bienveillance (*eunoia*). Ils combinent ainsi des qualités émotionnelles (proximité et bonté à l'égard des pauvres) et rationnelles (excellence de leur parcours scientifique, hauts standards d'exactitude et de rigueur). Cela renforce l'autorité de leurs propos, tout en les rendant, *via* les anecdotes personnelles, sympathiques et familiers au lecteur.

Cette analyse textuelle de *Poor Economics* démontre que Banerjee et Duflo ont réussi à amalgamer très efficacement des éléments souvent considérés comme antagoniques, y compris par les auteurs eux-mêmes : objectivité et subjectivité, abstraction et personnification, chiffres et narration, rationalité et émotion. Ils font appel à des figures de style comme à des composantes manipulatoires. Cette rhétorique efficace a « fasciné et convaincu » un « Nobel » comme Solow[16], dont la posture macroéconomique est pourtant aux antipodes de celle des auteurs. Elle ne doit pas occulter le manque d'épaisseur des récits, l'ampleur des angles

16. http://www.pooreconomics.com/about-book/what-others-are-saying.

morts des RCT et le danger de n'avoir que celles-ci « au menu » (Ravallion, chap. 1, ce volume).

La capacité à réunir différents publics autour d'un contenu commun

Les procédés rhétoriques à l'œuvre dans *Poor Economics* ont une autre caractéristique importante : ils « parlent » à des publics très divers, depuis un prix Nobel comme Solow jusqu'aux ONG, en passant par les décideurs politiques et le « grand public ». Le J-PAL ne tient pas un double ou un triple discours selon les publics auxquels il s'adresse. Les messages sont largement comparables, quels que soient les publics, des cours en ligne du J-PAL aux *Policy Briefcases*, de *Poor Economics* aux articles dans des revues du *top five*. Bien entendu, la forme du discours est fortement modulée pour chacun des publics : la technique statistique va être centrale dans l'article académique et rare dans l'article de vulgarisation. Néanmoins, la ligne argumentative reste largement la même.

C'est là une différence cruciale par rapport à d'autres courants économiques. Ainsi, des néo-keynésiens comme Krugman et Stiglitz recourent à des discours largement accessibles dans leurs ouvrages de vulgarisation scientifique et dans leurs blogs, des discours qui revêtent des dimensions interdisciplinaires et politico-économiques. C'est moins vrai de leurs productions académiques : elles relèvent essentiellement d'une modélisation en termes de théorie standard étendue (introduction d'imperfections de marché). Il existe chez eux des différences de fond dans le contenu du discours selon les publics. Prenons aussi l'exemple de Debreu, promoteur d'une « théorie économique sur le mode mathématique » (à savoir les mathématiques topologiques). Après avoir reçu son prix Nobel en 1983, il fut bien en peine de discourir sur l'économie réelle lorsque les journalistes le pressèrent de le faire[17]. Le discours de vulgarisation était pour lui informulable. Symétriquement, le grand public – comme les décideurs publics – ne pouvait accéder à ses publications écrites dans un langage mathématique hermétique. *Poor Economics* parvient ainsi à dépasser remarquablement les oppositions entre mondes savant et ordinaire, oppositions pourtant très fortes dans beaucoup de discours économiques. Cet avantage rhétorique comparatif constitue un élément déterminant du succès du J-PAL pour attirer à lui étudiants, chercheurs, journalistes et bailleurs de fonds. C'est un élément central de son modèle économique en pleine expansion. Jusqu'à présent, ce facteur rhétorique n'avait pas été mis au jour par les analyses du succès disciplinaire des RCT qui éclairent d'autres facteurs importants de l'économie politique des RCT (Labrousse, 2010 ; Jatteau, 2016 ; Bédecarrats *et al.*, 2019b).

17. Cet épisode m'a été rapporté par Alain Desrosières, qui le tenait lui-même d'une des filles de Debreu.

Des narrations pauvres

Les auteurs font un usage ubiquitaire et discrétionnaire de petites histoires malgré leur rejet explicite des anecdotes. Cela devient moins paradoxal si l'on considère leurs fonctions, éthique, de marketing social, de didactique et de persuasion. Ces petits récits de vie ont en outre une fonction heuristique discrète, mais cruciale pour faire sens des résultats expérimentaux et ouvrir la boîte noire expérimentale. Néanmoins, ces petites histoires relèvent de narrations minces, ancillaires et peu rigoureuses. Pour remplir pleinement cette fonction heuristique, elles nécessiteraient d'être enrichies par des narrations denses et rigoureuses dont les règles ont été explicitées en économie par DUMEZ et JEUNEMAÎTRE (2005). Il s'agirait de conjuguer non pas marginalement, mais explicitement approches quantitatives et qualitatives, notamment ethnographiques (MORVANT-ROUX *et al.*, 2014). Ainsi pourraient être développés des plans d'expérience et des interprétations plus pertinentes pour les environnements matériels, sociaux et culturels des sociétés dans lesquelles les expérimentations se déroulent.

Mais une telle évolution vers des méthodes mixtes a peu de chance de prendre place autrement qu'aux marges de l'économie. Les études rhétoriques montrent l'importance de l'*audience* dans la détermination de la forme et du contenu des discours. Or l'audience la plus fondamentale du J-PAL, celle à laquelle il mesure sa productivité (le nombre d'unités publiables par expérimentation, JATTEAU, 2016), celle qui produit des carrières, est celle des revues économiques les mieux cotées. Or ces dernières valorisent presque exclusivement la quantification.

Remerciements

L'auteure tient à remercier Thierry Guilbert, Stéphane Longuet, Robert Picciotto, Jonathan Morduch et les éditeurs pour leurs précieuses suggestions. Les insuffisances de l'article restent de l'entière responsabilité de l'auteure.

Chapitre 9

Les *randomistas* sont-ils des évaluateurs ?

Robert Picciotto

Introduction

Les *randomistas* prévoient un brillant avenir pour la théorie et la pratique du développement grâce à l'accumulation patiente de preuves expérimentales issues d'interventions menées au niveau individuel. Pour la charismatique cofondatrice du Laboratoire d'action contre la pauvreté (J-PAL) du Massachussets Institute of Technology (MIT), Esther Duflo, lauréate du prix Nobel 2019, une nouvelle ère de progrès scientifique s'annonce dans le champ social. Elle a ainsi fait sensation en déclarant, lors d'une conférence de la Banque mondiale sur l'évaluation de l'efficacité du développement : « La création d'une culture qui promouvrait, encouragerait et financerait des évaluations randomisées rigoureuses, et qui pourrait révolutionner les politiques sociales du XXIe siècle, tout comme les essais cliniques ont révolutionné la médecine du XXe siècle » (Lancet, 2004).

S'agit-il d'une mission réaliste pour les évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) ou d'une manifestation de la pensée magique ? Depuis le début du siècle, l'usage des RCT a massivement augmenté dans le domaine du développement : en un temps relativement court, elles se sont imposées sur un créneau très en vogue de la recherche en sciences sociales, celui de l'évaluation d'impact sur le développement. La publication annuelle d'évaluations d'impact expérimentales et quasi expérimentales connaît une véritable envolée. Elle plafonne aujourd'hui au niveau du pic qu'elle avait atteint en 2012, soit 400 à 500 études par an, ce qui est remarquable. Sur les 4 600 évaluations expérimentales et quasi expérimentales publiées, selon un recensement établi en juin 2018, seules 132 étaient antérieures à l'an 2000 (Cameron *et al.*, 2016).

Environ 62 % des évaluations d'impact figurant dans le référentiel de l'International Initiative for Impact Evaluation (3ie) sont exclusivement expérimentales et 5 % combinent RCT et méthodes quasi expérimentales. Le reste, soit un tiers

environ, repose exclusivement sur des méthodes quasi expérimentales. Certes, les RCT représentent encore à ce jour moins de la moitié des articles dans les revues économiques d'intérêt général, et moins d'un tiers de ceux présents dans les cinq principales revues dédiées à l'économie du développement (McKenzie, 2016). La hausse du nombre d'articles sur l'économie du développement publiés dans ces revues entre 1990 et 2015 est toutefois imputable pour les deux tiers aux RCT (Banerjee *et al.*, 2016). Comment expliquer alors cette progression rapide des RCT et que laisse présager leur adoption enthousiaste par les universités d'élite, les fondations philanthropiques et les instances de l'aide au développement pour l'avenir du mouvement en faveur de l'évaluation ?

Dans ce chapitre, je commence par relever la forte emprise que les RCT exercent sur l'imaginaire collectif, qui s'explique par les racines historiques profondes du courant expérimentaliste. Dans un deuxième temps, je montre que l'affirmation largement répandue selon laquelle les RCT constituent l'étalon-or de l'évaluation va à l'encontre du consensus acquis de haute lutte dans la communauté des évaluateurs : la diversité méthodologique constitue une bonne pratique. Troisièmement, j'observe que, malgré leurs limites, les RCT ont la faveur de ceux qui paient les exécutants de l'évaluation et qui dominent le marché actuel de l'évaluation. Quatrièmement, je reconnais que les RCT apportent une modeste contribution à la recherche en sciences sociales. Cinquièmement, et avant de conclure, j'établis que les RCT ont beau faire partie intégrante de la boîte à outils de l'évaluation, elles n'en sont pas pour autant des évaluations.

Éluder les dures leçons de l'histoire de l'évaluation

Les racines historiques du mouvement expérimentaliste sont profondes. Thalès de Milet, né au milieu des années 620 av. J.-C., a été le premier à proposer une interprétation des phénomènes naturels basée sur la théorie, s'écartant alors des explications surnaturelles ou mythologiques. Après lui, Platon et Aristote ont mis en avant des approches systématiques de l'étude de la nature par le biais du raisonnement déductif. Mais l'institutionnalisation de la recherche scientifique ne s'est imposée en Europe qu'au début de l'ère moderne.

Un engagement fondé sur la foi

Au fur et à mesure que l'expérimentalisme devenait une partie intégrante de la méthode scientifique, il a suscité un grand nombre de controverses et n'a acquis sa légitimité publique qu'en étant reconnu comme un renouveau de la religion innocente. L'appel à la sanction divine a été mobilisé pour valider le principe de base de la méthode scientifique, selon lequel la vérification expérimentale

est le seul test authentique du savoir. Par leur réexamen systématique des textes bibliques, John Milton et ses disciples ont proposé de nouvelles interprétations convaincantes de la Création. En fin de compte, leur conception réformiste de la foi religieuse a donné de la respectabilité au mouvement expérimentaliste (Picciotto, 2011).

Il s'est ensuivi une reconfiguration fondamentale de la relation entre la religion, la science expérimentale et la sphère publique. Pour Francis Bacon et ses disciples de la Royal Society, la pratique d'une observation et d'une évaluation minutieuses non corrompues par le dogme a été légitimée par l'avènement d'une nouvelle souche de l'apologétique chrétienne, qui a enjoint au public, ainsi qu'aux scientifiques et aux savants, d'apporter des preuves de la sagesse divine par l'examen direct de l'ordre naturel. Le positivisme a finalement étendu l'approche expérimentale à la société humaine en affirmant que, pour les sciences sociales comme pour les sciences physiques, seules sont valables les connaissances qui sont vérifiables, cumulatives, transculturelles et indépendantes de l'observateur.

La foi en l'expérimentation est ainsi devenue une composante de la doctrine religieuse, jusqu'à ce que la modernité émerge et que le désenchantement du monde prenne le dessus (Weber, 1958). Dès lors, l'expérimentalisme a été accepté tel qu'il était, sans référence à une quelconque divinité, mais ses caractéristiques sacrées ont persisté dans l'esprit collectif. Elles ont en effet été consacrées par Auguste Comte, le fondateur de la sociologie : sa « religion de l'humanité » fut inspirée par des principes positivistes. Depuis lors, la foi collective inébranlable en la supériorité de l'approche expérimentale a résisté à tout, même si ses hypothèses philosophiques fondamentales ont été discréditées.

Des fondements philosophiques fragiles

Aujourd'hui, la position épistémologique privilégiée par les partisans des RCT a été catégoriquement réfutée par les spécialistes des sciences sociales. Ils ne souscrivent plus au principe du positivisme logique, selon lequel il serait possible de formuler des généralisations immuables sur les relations humaines en dehors de tout contexte culturel spécifique. Ainsi, Émile Durkheim a d'abord soutenu que la sociologie avait pour mission de créer sa propre approche distinctive, plutôt que de reproduire les méthodes des sciences naturelles.

Max Weber s'est éloigné plus encore d'un positivisme étroit en suggérant que la complexité des interactions humaines est telle que les sciences sociales peuvent seulement révéler des relations causales au travers de simplifications hypothétiques des phénomènes sociaux. Le fossé entre les sciences sociales et les sciences naturelles s'est progressivement creusé sous l'impulsion de théoriciens critiques et de matérialistes historiques comme Karl Marx, Theodor Adorno et Jürgen Habermas. Leurs théories concurrentes ont convergé vers la proposition selon laquelle sciences naturelles et sciences sociales sont ontologiquement distinctes.

Thomas Kuhn a ensuite fait valoir que le choix de la théorie dans le domaine de la science est subordonné à des considérations paradigmatiques qui vont bien

au-delà de l'observation. Les critiques postmodernes sont allées plus loin et ont tenté de réfuter totalement la méthode scientifique en défendant l'idée que toute expérimentation est subjective, si ce n'est rétrograde, en particulier lorsqu'elle concerne la société. Ce plaidoyer frôlant l'irrationalité a fatalement exposé les déconstructionnistes à de vives critiques et à des accusations de subjectivité et de partialité. Mais le profond scepticisme suscité par des prétentions évaluatives qui n'explicitent pas leur finalité sociale s'était alors généralisé et le positivisme, en particulier dans sa forme utopique, avait perdu de sa superbe.

La science n'est plus considérée comme l'arbitre ultime de la politique sociale, et la croyance en un progrès humain – qui serait inévitablement alimenté par le développement technologique – n'a plus cours. En explorant l'interface entre le pouvoir et le savoir, la recherche sociale axée sur l'agir communicationnel dans la sphère publique est devenue un moyen privilégié d'exploiter l'évaluation pour promouvoir le bien collectif. Mais l'idée qu'il existe une réalité unique susceptible d'être identifiée de manière probante par l'observation, même en l'absence de théorie, a perdu tout crédit. Ainsi, Karl Popper a démontré que, dans le monde naturel comme dans le monde social, toute recherche scientifique est façonnée par les hypothèses avancées par les chercheurs, et que toutes les théories ne sont que de simples conjectures sujettes à réfutation : si la réalité existe, elle n'est vécue qu'indirectement et imparfaitement.

Ceci étant, la croyance constructiviste selon laquelle la réalité est une pure construction sociale demeure une position philosophique marginale. Un large consensus soutient que, si les expérimentations sont essentielles au progrès scientifique, la seule inférence valable qu'on puisse en tirer est la réfutation de théories prédéfinies de causalité. Sous cet angle, toute prise de décision rationnelle dans la sphère publique ne peut être guidée que par des connaissances contextuelles plausibles, bien que faillibles, issues d'une confrontation rigoureuse avec la réalité, d'une autocritique scrupuleuse, d'une critique par des pairs et d'un débat de principes.

Une loyauté à toute épreuve

Si le positivisme logique a perdu de son lustre dans les cercles philosophiques, il suscite encore une forte loyauté dans le monde universitaire. Les partisans des RCT estiment ainsi que les modèles expérimentaux constituent la *seule* base scientifique permettant de déterminer la causalité ou l'attribution. Cette position extrême est intenable, puisque la biologie, la géologie, l'astronomie, l'épidémiologie, les sciences médico-légales, etc. confirment toutes que la causalité peut être établie sans évaluations randomisées. Pour citer Lant Pritchett, « si l'expérimentation était la marque de fabrique de la science, il y aurait des prix Nobel d'alchimie, et non d'astrophysique[1] ».

Une observation et une évaluation minutieuses peuvent confirmer ou infirmer une théorie sur le monde naturel sans randomisation. La prédiction de la déviation

1. Communication personnelle.

de la lumière induite par la théorie générale de la relativité a été confirmée pour la première fois par Arthur Stanley Eddington à partir de ses observations de l'éclipse solaire du 29 mai 1919. Plus récemment, des essais utilisant des mesures radio interférométriques sur des quasars passant derrière le soleil ont confirmé la théorie de façon plus précise et cohérente.

De même, les RCT ne sont pas utiles dans l'administration de la justice. Les techniques d'enquête, les mécanismes de contestation et les règles de preuve sont jugés suffisants pour pénaliser, emprisonner et, dans certaines juridictions, exécuter des coupables présumés d'un crime. Les modèles randomisés ne sont pas non plus suffisamment flexibles pour prendre en compte la diversité des questions qui préoccupent les chercheurs en sciences sociales, la variabilité des contextes opérationnels ou la complexité des interventions de développement. Les approches qualitatives sont essentielles dans la quête de réponses aux dilemmes et aux défis du développement.

Mais les *randomistas* sont de vrais croyants. Ils excluent toute autre perspective, préfèrent s'associer avec d'autres croyants et tentent de vaincre la résistance des non-croyants en les excluant. L'une des caractéristiques distinctives du fondamentalisme est que la source de la vérité légitime réside dans le passé : les fondamentalistes se réfèrent souvent à des textes et à des personnages sacrés. De la même manière, les partisans radicaux des RCT bâtissent leur autorité sur les contributions intellectuelles des pionniers de l'évaluation, sans tenir compte des leçons apprises au cours du processus d'évolution de la discipline de l'évaluation.

Selon ALKIN (2004), toutes les doctrines d'évaluation actuelles peuvent être classées selon l'importance qu'elles accordent aux méthodes, aux usages ou à la valorisation. Il propose une métaphore où la théorie de l'évaluation est un arbre touffu, composé de modèles d'évaluation concurrents, qui se regroupent en trois branches principales. Le modèle expérimentaliste occupe une place prépondérante à la base même de la branche méthodologique, car il est présent depuis la création de la discipline de l'évaluation.

Des conceptions évolutives de l'évaluation

De façon plus spécifique, les pionniers de l'évaluation des programmes sociaux ont conçu l'évaluation comme une courroie de transmission entre les sciences sociales et les décideurs[2]. Ainsi, Donald T. Campbell, le méthodologue de la *Société expérimentale*, a présenté les interventions publiques comme des expérimentations politiques. Très concentré sur l'élimination des biais dans la recherche en sciences sociales, il a vanté l'expérimentation comme « le seul moyen de régler les différends relatifs aux pratiques d'enseignement, la seule façon de vérifier les améliorations en matière d'enseignement et le seul moyen de fonder une tradition cumulative » (CAMPBELL et STANLEY, 1963 : 2).

2. L'avènement de la discipline de l'évaluation coïncide également avec le début des efforts de développement, une période d'optimisme où les épées de la Seconde Guerre mondiale ont été transformées en socs de charrue par les alliés victorieux.

Les *randomistas* continuent de soutenir ce point de vue, bien que Campbell ait fini par reconsidérer et nuancer sa position méthodologique. En effet, au vu des résultats décevants des études expérimentales dans le domaine des politiques sociales, il a révisé son avis négatif sur les méthodes qualitatives. Il a admis qu'un jugement qualitatif expert était nécessaire pour identifier ou écarter de potentielles relations de causalité, ou encore pour interpréter les effets secondaires des interventions publiques. Ainsi, afin de « d'être vraiment scientifique, on doit rétablir le fondement qualitatif du quantitatif » (CAMPBELL, 1974).

Thomas Cook s'est inspiré des idées de Campbell en se concentrant sur les facteurs contextuels et leur incidence sur les expérimentations classiques. Il a mis au point des techniques quasi expérimentales permettant de surmonter les difficultés liées au contrôle de l'expérimentation. Il a également souligné l'importance de la concertation avec les parties prenantes de l'évaluation. De même, Peter Rossi et Carol Weiss, tout en reconnaissant l'attrait des expérimentations contrôlées pour éliminer les biais de sélection, ont apporté des contributions majeures au champ méthodologique en corrélant la logique d'intervention sous-jacente aux programmes publics avec les évaluations fondées sur la théorie.

Le parcours intellectuel de Lee J. Cronbach l'a éloigné d'une adhésion systématique aux essais randomisés pour aboutir à son rejet total de l'expérimentalisme classique. Cronbach est arrivé à la conclusion que seules les décisions simplistes « *go/no go* » sont influencées par les essais randomisés, alors que la collecte de données d'évaluation utiles pour l'action nécessite d'explorer un large éventail de questions pertinentes plutôt que de se focaliser sur la série de questions nécessairement restreinte qui se prête à des évaluations randomisées.

L'intérêt de Cronbach pour l'élaboration de politiques éclairées par l'évaluation l'a finalement amené à remettre en question la validité externe des évaluations randomisées. Il a fini par douter du fait que des généralisations robustes sur le comportement humain puissent être établies par le biais de la recherche en sciences sociales, et a prôné davantage de modestie et de retenue dans la formulation des recommandations politiques (CRONBACH, 1982). De la même façon, STAKE (2010), qui a commencé sa carrière d'évaluateur comme positiviste et mathématicien, s'est montré de plus en plus désabusé par le potentiel de la mesure et de la modélisation formelle pour l'évaluation des programmes sociaux.

Retour vers le futur ?

L'ignorance réelle ou feinte de l'histoire de l'évaluation a condamné l'industrie du développement à la répéter. Les débats fomentés par les vrais apôtres des RCT ne sont pas nouveaux. Les conflits acharnés entre partisans des méthodes quantitatives et qualitatives ont longtemps fracturé le monde de l'évaluation et de la recherche en sciences sociales, jusqu'à ce qu'ils trouvent une issue dans les années 1990. Après d'innombrables débats et à la lumière de multiples publications, presque une décennie avant que le MIT ne crée son Laboratoire d'action contre la pauvreté, les « guerres de paradigmes » ont été résolues avec succès à la satisfaction de la plupart des chercheurs et évaluateurs en sciences

sociales : qu'elles soient qualitatives et quantitatives, les deux types de méthodes ont leur utilité, et les méthodes mixtes ont l'avantage (DATTA, 1994).

Le retour vers le futur des RCT dans le développement international depuis le début du siècle est donc paradoxal. Les experts s'accordent à dire qu'une diversité méthodologique adaptée au contexte l'emporte sur l'adhésion rigide à un modèle d'évaluation unique. Les partisans des RCT ont choisi d'ignorer ce consensus acquis de haute lutte. Ils restent bornés au constructionnisme utopique des pionniers de l'évaluation. Ils affirment que les expérimentations de terrain présentent des avantages uniques pour produire des essais rigoureux sur l'efficacité de l'aide et générer des connaissances scientifiques sur le développement. Peu importe que les responsables des politiques de développement apprécient depuis longtemps la capacité démontrée de l'évaluation indépendante utilisant des méthodes mixtes à promouvoir l'autoévaluation, à suivre les performances, à tirer parti des leçons de l'expérience et à reconsidérer les approches erronées des politiques de développement (GRASSO *et al.*, 2003). Les RCT occupent une place centrale dans la recherche économique.

Ainsi, le dédain envers l'histoire de l'évaluation et les objections doctrinales aux évaluations qualitatives sous-tendent la popularité croissante des RCT dans le développement. L'échec de la recherche sociale macro-économique à satisfaire les sceptiques de l'aide a facilité l'incursion des micro-économistes sur le marché de l'économie du développement. La montée en puissance des RCT était typiquement liée à la désillusion quant à la capacité des méthodes macro-économiques à générer des prescriptions politiques valables pour le secteur de l'aide : une industrie artisanale d'études de recherche politique fondée sur des régressions multi-pays avait en effet généré des conclusions diverses et contradictoires sur l'impact global de l'aide (TARP, 2009).

La recherche macro-économique n'a pas su identifier les corrélations solides qui existent entre les volumes d'aide, les prescriptions politiques et la croissance économique visée par les décideurs politiques. Et ce n'est guère surprenant : parfois, l'aide fonctionne, parfois elle échoue. Le contexte est important et les objectifs de l'aide varient. Le développement n'est pas qu'une question de croissance. Les macro-modèles peuvent difficilement cerner les apports de l'aide sur le plan de la technologie et du renforcement des capacités. Les circuits, les instruments et les modalités de l'aide sont importants. Et cela est vrai aussi des contextes sociaux et institutionnels.

Pourtant, à une époque où l'*establishment* de l'aide est en plein bouleversement, les résultats ambigus ont contribué à décourager le public quant à l'utilité de la recherche macro-économique pour déterminer l'impact de l'aide. Ainsi, un ensemble d'articles tendant à expliquer les écarts entre l'impact de l'aide à l'échelle nationale et les études au niveau des projets (le fameux « paradoxe micro-macro ») a par exemple été publié, semant un doute supplémentaire parmi les chercheurs sur le bilan des résultats de l'aide publiés par les services d'évaluation des agences de développement. Les conclusions sur les performances de développement, même si elles sont basées sur des méthodes qualitatives

transparentes, ont soudain été jugées peu fiables par les expérimentalistes, qui considèrent que seules les méthodes quantitatives constituent des tests d'attribution valides.

Les micro-économistes entrent dans la mêlée sur l'efficacité de l'aide

Dans un environnement intellectuel tumultueux, deux factions en guerre – les optimistes de l'aide, conduits par le professeur SACHS (2005) de l'université de Columbia et les pessimistes de l'aide, inspirés par EASTERLY (2007) de l'université de New York – se sont engagées dans des joutes intellectuelles qui ont échauffé les esprits plus qu'elles n'ont éclairé les débats, sapant ainsi la confiance collective dans l'aide au développement et offrant une opportunité stratégique aux jeunes économistes du MIT. Les résultats forcément peu concluants des recherches en sciences sociales ont déplacé le point de mire du débat sur l'efficacité de l'aide depuis le plan abstrait de la macro-économie vers le terrain de jeu plus concret de la micro-économie.

Se gardant bien de généralisations grandiloquentes, les *randomistas* ont prôné une nouvelle approche axée sur l'examen clinique d'interventions de développement spécifiques. Peu importe que les évaluateurs du développement aient toujours cherché à vérifier si les hypothèses des praticiens de l'aide « fonctionnaient sur le terrain » aux niveaux des projets, des secteurs et des pays. Leurs travaux ont été rejetés sans appel par ceux qui postulent que seules les méthodes expérimentales sont valables, alors même qu'il est amplement prouvé que les évaluations qualitatives du développement ont longtemps été, et demeurent, des instruments essentiels pour assurer un suivi approprié dans l'administration de l'aide.

Le flou des résultats de recherche au niveau macro-politique, conjugué aux critiques superficielles des évaluations qualitatives, a été renforcé par le scepticisme à l'égard de l'aide. Étonnamment, les preuves abondantes apportées par les évaluations de projets et le succès extraordinaire des efforts de développement dans de nombreuses économies émergentes ont été considérés comme sans importance, au motif douteux que l'attribution ne peut être établie sans expérimentation.

L'expérimentalisme nouvelle formule

La combinaison d'une rigueur scientifique présumée, d'une neutralité idéologique appliquée et d'un pragmatisme volontariste s'est avérée irrésistible. Elle a rapidement gagné le soutien enthousiaste de fondations philanthropiques internationales désireuses de s'imposer sur la scène du développement. Avec l'aide financière de la fondation Bill & Melinda Gates et de la fondation William et Flora Hewlett, un groupe de travail sur les lacunes en matière d'évaluation (*Evaluation Gap Working Group*) a été constitué par le Center for Global Development (CGD) en 2004. Sa raison d'être sous-jacente était que des milliards de dollars et des milliers de programmes d'aide avaient été dédiés à la santé, à l'éducation et à

d'autres domaines du secteur social sans qu'aucune étude ne puisse déterminer sans ambiguïté si ces programmes « fonctionnaient » réellement.

Le rapport du groupe de travail (Center for Global Development, 2006) a rejeté le système de notation utilisé par les évaluateurs du développement pour mesurer l'efficacité des interventions d'aide. Il a conclu que les résultats des évaluations traditionnelles manquaient de validité, car celles-ci n'abordaient pas la question de l'attribution de manière rigoureuse. Une recherche systématique de preuves solides sur l'efficacité des interventions de développement par des méthodes « scientifiques » a été préconisée. Le rapport considérait que c'était la seule façon d'obtenir des preuves adéquates pour mettre fin aux programmes inefficaces et identifier les approches de réduction de la pauvreté qui méritaient d'être répliquées[3].

Le rapport soutenait en particulier que déterminer si « l'aide marche », requiert des expérimentations ou des méthodes quasi expérimentales se rapprochant de l'étalon-or de la randomisation[4]. Or, comme indiqué plus haut, ce statut d'étalon-or avait été mis à mal des décennies auparavant. De toute évidence, les enseignements de la « guerre des paradigmes » n'avaient pas été intégrés par le petit monde de l'économie du développement, et rien n'a pu arrêter la dynamique enclenchée par les arrivistes du MIT. Progressivement, le financement de la recherche pour le développement s'est détourné des études macro-économiques au profit des évaluations micro-économiques des interventions de développement.

De la recherche en sciences sociales à l'évaluation

Il n'a pas fallu attendre longtemps avant que la lutte pour la suprématie des RCT au sein de l'élite de la recherche en sciences sociales ne gagne le monde de l'évaluation et ne ravive le conflit latent des paradigmes. Les évaluateurs de l'aide qui venaient seulement de rejoindre le courant dominant de la profession ont été pris au dépourvu. Non préparés à l'assaut, ils ont cédé du terrain. Ils n'avaient pas pris part aux débats méthodologiques qui avaient secoué la communauté de l'évaluation à la fin des années 1970 et au début des années 1980[5]. C'est ainsi que les micro-économistes inféodés aux méthodes expérimentales ont envahi un territoire jusque-là réservé aux praticiens du développement. De bruyantes controverses ont rapidement éclaté lors de conférences internationales, et un schisme de la communauté de l'évaluation du développement s'est avéré inévitable.

À une extrémité du spectre, des évaluateurs du développement chevronnés, formés aux méthodes qualitatives, ont jugé illusoire la rigueur attribuée aux

3. Dans la pratique, cette vision ne s'est jamais concrétisée.

4. En médecine, un étalon-or désigne une méthode de diagnostic ou de comparaison considérée comme irréfutable.

5. Le conflit méthodologique dans le monde de l'évaluation a de nouveau brièvement éclaté fin 2003 aux États-Unis, lorsque le ministère de l'Éducation a décidé de privilégier les méthodes expérimentales dans le financement des évaluations.

méthodes expérimentales. À l'autre extrémité, ceux qui avaient longtemps cherché à se rapprocher de l'économie, « reine des sciences sociales », ont vu d'un bon œil les incursions des micro-économistes dans le domaine de l'évaluation et ont plaidé en faveur d'une collaboration étroite. Après de longues délibérations, ils se sont mis d'accord sur un document d'orientation méthodologique (LEEUW et VAESSEN, 2009), qui reconnaissait la supériorité fréquente des modèles expérimentaux pour établir l'attribution, mais réfutait l'hypothèse selon laquelle les évaluations randomisées constituaient un étalon-or. Il préconisait plutôt des méthodes mixtes adaptées aux besoins spécifiques de chaque évaluation.

Ce « jugement de Salomon » marquait une nouvelle trêve. Mais le consensus ne s'est guère étendu au-delà des cercles appartenant au courant dominant de l'évaluation. En revanche, dans le monde de la recherche en sciences sociales, et pour les principaux utilisateurs des évaluations, les malentendus et les tensions persistèrent. De toute évidence, le conflit a été mis en suspens. Quel est donc le consensus des experts concernant les RCT ?

Le potentiel et les limites des méthodes expérimentales

Dans des circonstances adaptées et entre des mains expertes, les méthodes expérimentales fournissent une estimation des résultats qui auraient été observés si l'intervention n'avait pas eu lieu. Pour cela, elles cherchent à établir une comparabilité stricte entre les groupes de contrôle et de traitement en sélectionnant de façon aléatoire des bénéficiaires et des non-bénéficiaires issus d'une même population par un processus reposant explicitement sur le hasard (par exemple, un lancé de dés, des tours de roulette ou une table de nombres aléatoires).

Une répartition non biaisée signifie que la probabilité de se retrouver dans le groupe de contrôle ou le groupe de traitement est identique. Cette caractéristique des RCT vise à résoudre le problème du *biais de sélection*, qui se produit lorsqu'une comparaison des impacts sur deux ensembles très différents de bénéficiaires finit par attribuer à tort les résultats observés à l'intervention, alors que différentes caractéristiques connues ou inconnues des groupes avec et sans traitement peuvent avoir opéré.

C'est par exemple souvent le cas lorsque les personnes qui ont accès au programme sont plus riches, plus puissantes, plus motivées ou plus instruites. En principe, une véritable assignation aléatoire aux groupes de traitement et de non-traitement issus d'une même population permet de garantir que, exception faite des fluctuations dues au hasard, l'impact de l'intervention peut être déterminé de manière fiable en comparant les résultats entre les deux groupes et en veillant à ce que tous les autres facteurs susceptibles d'altérer les résultats soient identiques, sauf erreurs stochastiques.

Afin de vérifier la fiabilité des essais, des techniques statistiques sont disponibles pour déterminer l'intervalle de confiance que l'on peut sans risque attribuer au résultat (c'est-à-dire le rôle qu'a pu jouer le hasard pur associé au processus de randomisation). Les RCT présentent ainsi l'avantage supplémentaire de permettre aux évaluateurs d'établir une mesure de la significativité statistique des résultats de l'évaluation.

Les limites des RCT

Comme l'explique clairement le chapitre de Ravallion (chap. 1, ce volume), il serait erroné de prétendre que toute différence entre les résultats du groupe de traitement et du groupe de comparaison ne peut être due qu'à l'intervention. En effet, ce n'est que si le groupe de traitement, le groupe de contrôle et le processus qui affecte chacun d'eux sont strictement identiques (sauf en termes de cause et d'effet) que des conclusions fiables peuvent être tirées. Pourtant, les erreurs d'échantillonnage sont inévitables et la validité interne peut être compromise par des facteurs latents et non observés qui n'ont pas été pris en compte lors de l'élaboration des groupes de traitement et de contrôle.

Ces écueils statistiques ne sont pas souvent reconnus par les défenseurs des RCT, et il n'est pas toujours possible de les surmonter à un coût raisonnable. Quelle est alors l'applicabilité des évaluations randomisées pour mesurer l'impact des interventions de développement ? Elles peuvent être utiles si leurs risques sont identifiés et traités. Cela étant, elles ne sont pas toujours appropriées. Elles ne se concentrent que sur un seul paramètre de la politique, alors que la plupart des interventions de développement sont motivées par des théories d'action et de changement complexes et visent de nombreux objectifs politiques. Elles supposent également que les interventions sont fixes et stables alors que, dans le monde réel, elles sont flexibles et adaptables.

Les RCT sont redondantes lorsqu'il n'existe aucune autre explication plausible concernant les résultats observés. Elles ne sont pas toujours une option réalisable. Il n'est par exemple pas possible de randomiser l'emplacement géographique de projets d'infrastructure (Ravallion, 2009a). Les méthodes expérimentales ne sont pas applicables lorsqu'aucun groupe cible non traité ne peut être identifié, par exemple lorsqu'une intervention a une portée universelle (imposition d'une limite légale pour la consommation d'alcool, programme de réforme de la fonction publique, libéralisation d'un régime d'importation, etc.) ou lorsque le modèle de l'intervention est flexible et adaptable à des circonstances nouvelles (Lensink, 2014).

La validité externe n'est pas non plus le point fort des méthodes expérimentales. Même lorsque les expérimentations sont appropriées, elles peuvent ne pas répondre aux besoins des décideurs politiques, qui se préoccupent, avant tout, non pas de ce qui s'est passé dans un échantillon expérimental d'essai, mais des chances que cela continue à fonctionner dans un environnement de mise en œuvre diversifié, complexe et instable (Cartwright et Munro, 2010).

La taille, la structure et le contexte des programmes sont déterminants pour le résultat des activités de développement.

L'argument plaidant en faveur des études observationnelles et des études qualitatives réside également dans le fait que seules les expérimentations fondées sur une théorie plausible valent la peine d'être menées. Ainsi, les examens systématiques qui agrègent les conclusions d'études sur les transferts monétaires conditionnels sans tenir compte des élasticités différentielles de la demande sont quasiment dénués de sens. Pour obtenir une évaluation de qualité, il est essentiel de bien comprendre le fonctionnement d'un programme dans son contexte spécifique et de préciser la théorie sur laquelle se fondent les conclusions. La compréhension appropriée des relations causales et l'identification des explications concurrentes à réfuter requièrent une connaissance approfondie de l'intervention, de sa conception, de ses protocoles de mise en œuvre et des motivations des participants et des bénéficiaires du programme.

Même lorsque les expérimentations visant à définir l'attribution paraissent sensées, elles nécessitent d'excellentes compétences, des études poussées, de grands échantillons et des dispositifs spécifiques d'assurance qualité. Or, ces conditions préalables ne peuvent pas toujours être réunies dans le domaine du développement. De ce fait, les RCT sont susceptibles de ne pas permettre une utilisation parcimonieuse des rares ressources d'évaluation. Elles peuvent également empêcher le recours à des évaluations moins coûteuses et plus efficaces, et entraver la pleine participation des bénéficiaires de l'aide au processus d'évaluation en transférant le contrôle d'une évaluation d'impact économétriquement sophistiquée à des universités bien dotées et à des groupes d'experts situés dans des pays développés.

Préoccupations éthiques

Les RCT constituent une réponse au biais de sélection lorsque les personnes qui accèdent au programme sont plus riches, plus puissantes, plus motivées ou plus instruites. L'assignation aléatoire aux groupes de traitement et de non-traitement à partir de la même population garantit qu'à l'exception des fluctuations dues au hasard, l'impact de l'intervention peut être déterminé de manière fiable en veillant à ce que tous les facteurs susceptibles d'altérer les résultats soient identiques, sauf erreurs stochastiques. Les évaluations randomisées fournissent également aux évaluateurs une mesure de la significativité statistique des résultats de l'évaluation.

Ce sont là de formidables avantages. Mais les méthodes expérimentales soulèvent presque invariablement des préoccupations éthiques qui ne sont pas souvent reconnues par les *randomistas*. Il peut être discriminatoire, voire illégal, de priver les membres du groupe de contrôle d'un traitement utile sur la base d'un processus de sélection perçu comme capricieux et arbitraire. Dans certains territoires, il est interdit d'administrer aux membres du groupe de comparaison un traitement inférieur au meilleur traitement disponible.

De même, il n'est généralement pas considéré comme une pratique éthique d'inciter les membres d'un groupe de traitement à participer à une intervention qui peut avoir des effets secondaires négatifs. Paradoxalement, les procédures de consentement éclairé appliquées dans de tels cas peuvent introduire le biais de sélection que la méthode est justement censée éviter, de sorte que des expérimentations en aveugle doivent être utilisées. Même dans ce cas, on ne peut pas éliminer les effets subtils que les expérimentations peuvent induire sur les groupes de traitement et de non-traitement (biais de Hawthorne et de John Henry).

Effets imprévus

En privilégiant les interventions publiques qui peuvent être évaluées par des méthodes expérimentales, on encourage la sélection de programmes et de projets simplistes qui peuvent ne pas être adaptés à l'objectif visé et/ou favoriser l'évitement des questions évaluatives essentielles au profit des seules questions qui se prêtent à la randomisation. Les RCT ne peuvent pas répondre seules aux questions suivantes : « pourquoi ? », « qui ? » et « et alors ? ».

La plupart des politiques, programmes et projets de haut niveau qui sont privilégiés aujourd'hui par les agences internationales de développement ne peuvent être évalués par un traitement randomisé. Cela signifie que la randomisation convient surtout aux problématiques limitées ou aux projets simples, avec des participants et des non-participants facilement identifiables, et dont les externalités ne risquent pas de fausser les résultats. Elle est peu adaptée à l'évaluation de programmes complexes ou compliqués dans des environnements instables. Or, c'est justement là que le manque de connaissances est le plus important.

Il existe des alternatives

De nombreux évaluateurs mènent toute leur carrière sans recourir à la moindre évaluation randomisée. Cela s'explique en partie par le fait que d'autres méthodes sont mieux adaptées pour déterminer *pourquoi* les interventions réussissent, *si* des problèmes de conception ou de mise en œuvre expliquent les échecs constatés des interventions ou *qui*, parmi les partenaires du développement, est responsable des résultats observés. Elles impliquent la participation, l'observation, l'analyse d'informations textuelles, des réunions de village, des entretiens non directifs, etc.

Pour permettre une analyse économétrique, la collecte de données qualitatives nécessite bien entendu un codage minutieux et une quantification systématique. Les méthodes qualitatives guidées par des théories du changement étudient ce qui s'est passé et pourquoi. Elles sont mieux à même d'expliquer pourquoi les effets escomptés ont été atteints ou non (ainsi que l'étendue et la nature des effets imprévus). Elles aident à faire la distinction entre les problèmes liés à la conception et ceux relevant de la mise en œuvre.

Alors que les méthodes expérimentales sont fondées sur des données, les approches qualitatives basées sur la théorie sont définies par les questions qui intéressent les parties prenantes et par les hypothèses qui sont intégrées dans les interventions des programmes et des projets (BAMBERGER *et al.*, 2010). Enfin, il existe une grande variété d'outils permettant de simuler un contrefactuel, sans passer par la randomisation. La liste qui suit n'est qu'une indication de la richesse des méthodes et outils dont disposent les évaluateurs. Elle ne prétend nullement évaluer leurs forces et faiblesses respectives dans divers contextes d'évaluation.

Analyse de régression et analyse factorielle

L'analyse de régression permet de déterminer dans quelle mesure diverses caractéristiques relatives au contexte et aux bénéficiaires d'une intervention expliquent les variations des effets obtenus. Le résultat est imputable au programme, en supposant que toutes les explications concurrentes ont été prises en compte dans le modèle. La *régression par discontinuité* compare les effets du traitement sur des sujets sélectionnés selon un critère (par exemple, la notation des sujets par des experts en fonction de leur probabilité de réussite, ou de leur besoin de bénéficier de l'intervention). Elle compare l'effet du traitement juste au-dessus d'un seuil d'éligibilité avec ceux obtenus juste en dessous de ce seuil.

Modèles quasi expérimentaux

Lorsque la randomisation n'est pas réalisable, elle peut être simulée au moyen de modèles *quasi expérimentaux*. Les personnes incluses dans les groupes de traitement et de non-traitement sont *appariées* pour assurer une certaine similarité au niveau des caractéristiques susceptibles d'influencer le résultat. Des ajustements statistiques peuvent être pratiqués pour garantir une étroite ressemblance entre les deux groupes sur ces dimensions pertinentes.

Modélisation statistique multivariée

Conçu pour prendre en compte toutes les relations supposées entre les variables de traitement et de non-traitement, ce modèle doit permettre d'expliquer les différences entre les deux groupes au stade initial afin que les différences observées au stade post-traitement puissent être compensées statistiquement. Mais cette approche présente des inconvénients propres : elle suppose non seulement que le modèle a cerné avec précision les relations entre les variables, mais aussi que tous les facteurs expliquant les différences avant traitement ont été identifiés.

Approches participatives

L'évaluation d'impact qualitative repose sur les perceptions exprimées par les bénéficiaires réels ou potentiels, les observateurs experts et/ou les décideurs. Le vote par couleur facilite le débat de principe en affichant les opinions des parties prenantes par des présentations en couleurs de leurs votes (ou scores) sur des questions clairement formulées concernant l'intervention. La schématisation conceptuelle implique l'utilisation de tableaux à feuilles mobiles et de cartes (ou

de logiciels de traitement des données) pour obtenir une image graphique des perceptions des parties prenantes sur les impacts potentiels d'une intervention de développement. Elle fait appel à des modérateurs expérimentés pour motiver un groupe représentatif de parties prenantes qui sont bien informées et décidées à participer.

Enquêtes et échantillonnage

La collecte et l'interprétation des données d'enquête, les entretiens structurés ou semi-structurés, les groupes de discussion et autres méthodes d'implication des bénéficiaires peuvent permettre de comprendre ce qui fonctionne, ce qui ne fonctionne pas et pourquoi. Lorsque des groupes importants de citoyens ou de bénéficiaires sont interrogés, la collecte et l'interprétation des données exigent des stratégies d'échantillonnage efficaces.

Méthode d'élimination générale

SCRIVEN (2008) a proposé une alternative aux RCT inspirée par les techniques d'enquête criminelle et mettant l'accent sur les mobiles, les moyens et l'opportunité. Cette méthode d'élimination générale nécessite une revue des travaux publiés et/ou la consultation de personnes possédant une expertise tacite pertinente sur le domaine de l'intervention. Le processus commence par un recensement systématique des causes possibles relevant de l'intervention. Ensuite, une liste des *modus operandi* est établie pour chaque cause possible. Elle est suivie d'un examen détaillé des faits relatifs au cas. Seules les causes qui tiennent encore sont retenues comme explications potentielles.

Panels d'experts

Le recours à des panels d'experts composés de spécialistes indépendants qui connaissent bien le domaine de l'intervention peut être utile en combinaison avec d'autres méthodes, notamment lorsque l'équipe d'évaluation ne comporte pas de spécialistes du domaine ou d'évaluateurs chevronnés. Ces panels peuvent être utilisés pour déterminer si les impacts observés sont conformes à ce qui peut raisonnablement être attendu dans un contexte spécifique. La validité et la fiabilité des jugements des panels d'experts peuvent être améliorées grâce à la *méthode Delphi (ou méthode de Delphes)*, à savoir des procédures de consultation de chacun des experts sans concertation préalable entre eux.

Étude comparative (benchmarking)

L'étude comparative utilise des tests de performances clés pour juger de l'impact par des comparaisons avec les bonnes ou les meilleures pratiques observées dans des circonstances similaires. L'étude comparative interne identifie et cherche à reproduire les bonnes pratiques observées au sein d'un programme. L'étude comparative externe compare l'impact d'une intervention avec celui d'une initiative caractérisée par des conditions similaires et réputée avoir atteint des normes d'excellence.

Le marché actuel de l'évaluation favorise les RCT

Étant donné l'écrasant consensus auquel la communauté de l'évaluation est parvenue concernant les sérieuses limites des RCT, qu'est-ce qui explique l'ascension extraordinaire du mouvement expérimentaliste pour les évaluations en matière de développement international ? Quoi qu'en pensent les experts, en pratique, la politique d'évaluation tend naturellement à refléter les intérêts dominants dans la société. Par conséquent, les concepts d'évaluation les plus influents à un moment donné traduisent les modèles mentaux qui motivent les décisions des puissants de ce monde.

Les dynamiques qui en résultent sont appréhendées de façon très pertinente par VEDUNG (2010) dans son célèbre modèle de diffusion des évaluations. Il raconte l'histoire de l'évaluation comme une succession de vagues poussées par les vents changeants de l'idéologie politique. Chaque vague est entraînée par la marée des doctrines du moment. La vague finit par perdre de l'énergie et, une fois arrivée au bout de sa course, elle laisse derrière elle des couches de sédiments intellectuels qui enrichissent la discipline et en façonnent les contours.

Les vagues de diffusion de l'évaluation

Le mouvement expérimentaliste est emblématique de la première vague et, comme nous l'avons déjà mentionné, les hypothèses positivistes qui le sous-tendent ont progressivement été délaissées et, sous les gouvernements démocratiques des États-Unis, une vague constructiviste, participative et pluraliste *axée sur le dialogue* a déferlé à la fin des années 1960, lorsque les valeurs sous-jacentes de la guerre intérieure contre la pauvreté et l'aide internationale se sont rencontrées. Puis, dans les années 1980, le vent politique s'est mis à souffler brusquement vers la droite. Une troisième vague *néolibérale* puissante s'est alors formée, engloutissant la discipline de l'évaluation. Marquée par le nouveau courant de pensée du *New Public Management*, elle a supplanté les approches évaluatives constructivistes, dialogiques, participatives et démocratiques de la deuxième vague.

Nous surfons maintenant sur une quatrième vague. Elle *repose sur des preuves empiriques* et tient le néolibéralisme pour acquis. Elle est axée sur la réalisation des objectifs et privilégie les méthodes quantitatives. Elle légitime une évaluation dénuée de toute valeur en la parant d'attributs technocratiques. Elle accorde une place de choix à l'accomplissement des buts politiques fixés par les dirigeants. Elle prospère en surveillant les progrès au moyen d'indicateurs athéoriques. Dans cet environnement favorable, une approche technocratique, positiviste, axée sur l'utilisation et fortement dépendante des méthodes expérimentales répond tout à fait aux exigences d'un marché de l'évaluation de plus en plus dominé par des intérêts particuliers.

Paradoxalement, le même milieu intellectuel qui aspirait à des méthodes d'évaluation plus rigoureuses a fait naître de nouvelles menaces pour l'intégrité des processus d'évaluation et la validité de leurs résultats. Selon HOUSE (2014), « en raison des changements structurels intervenus dans la société, nous faisons face à un nouvel ensemble de biais potentiels, une famille de biais que nous devons ou devrions gérer ». Parmi ces changements structurels, on note l'envahissement progressif des affaires publiques par des intérêts privés exacerbés. Les travaux d'évaluation dans la recherche médicale (toujours acclamés et qualifiés d'exemplaires par les défenseurs des RCT) sont emblématiques des risques auxquels est actuellement confrontée l'entreprise d'évaluation du développement.

Les sirènes de la recherche médicale

Amorcer une initiative de transformation sociale par le biais d'une intervention de développement n'est pas la même chose que d'administrer un médicament. Cela ne signifie pas pour autant que le travail scientifique n'est pas capable de rigueur dans le domaine de la recherche médicale ou que la randomisation n'est pas une méthode de choix pour évaluer l'attribution dans certaines circonstances. Mais il faut avoir conscience des écueils de la recherche médicale telle qu'elle est actuellement pratiquée avant de la transposer à l'identique dans le domaine de l'évaluation du développement.

Dans la pratique, les études de recherche médicale évaluées par des pairs et diffusées par les médias grand public ont annoncé des conclusions différentes concernant les bienfaits sur la santé de traitements comme la prise régulière de vitamines, la prise d'une aspirine par jour, le fait de dormir plus de huit heures par nuit, la consommation de vin rouge à chaque repas, les risques de cancer liés à l'utilisation des téléphones portables, le fait d'habiter à proximité d'une ligne électrique à haute tension, etc. Des allégations extravagantes et parfois frauduleuses ont réussi à passer à travers les mailles du processus d'examen par des pairs des revues scientifiques. Ainsi, une grande évaluation randomisée a révélé que les prières secrètes d'inconnus peuvent sauver la vie de patients ayant subi une chirurgie cardiaque, alors qu'une autre a prouvé qu'elles peuvent leur nuire (FREEDMAN, 2010).

IOANNIDIS (2005a), directeur du Prevention Research Center de l'université de Stanford, a conçu un modèle mathématique permettant d'évaluer la probabilité qu'un résultat de recherche médicale soit vrai. Son article de référence confirme que la probabilité des hypothèses dépend de bien d'autres éléments que le seuil de l'intervalle de confiance fixé à 5 % par la plupart des revues. Ses simulations montrent en particulier qu'une mauvaise sélection de la relation testée, une puissance insuffisante des modèles statistiques, des traitements médicaux caractérisés par des effets mineurs, diverses sources de préjugés chez les chercheurs, etc. ont eu un effet dévastateur sur la validité de la plupart des résultats de recherche publiés.

Même à des niveaux modestes, les biais des chercheurs (alimentés par l'ambition ou la conviction) sont propices à une mauvaise interprétation des essais statistiques, à une utilisation faussée des preuves et/ou à une présentation

trompeuse des résultats. Ceux publiés sur la recherche médicale sont souvent manifestement faux. Et même les résultats les plus salués peuvent ne pas être dignes de confiance (IOANNIDIS, 2005b). Cette érosion de la crédibilité de la recherche médicale est due à l'invasion des intérêts particuliers, un risque qui pèse de plus en plus sur le monde de l'évaluation.

Jusqu'aux années 1980, la recherche sur les médicaments était largement indépendante des sociétés pharmaceutiques. Ce n'est plus le cas : les essais cliniques sont désormais contrôlés par des multinationales privées et les RCT ne protègent pas le processus contre les nombreux biais systémiques (HOUSE, 2008) :

– les nouveaux médicaments sont souvent testés en parallèle avec des placebos (le contrefactuel sélectionné), plutôt qu'avec des médicaments en usage, ce qui conduit souvent à recommander des variantes mineures de médicaments existants, même si elles ne sont pas supérieures à ceux-ci ;

– les comparaisons entre médicaments concurrents ne se basent pas toujours sur des dosages équivalents ;

– des sujets plus jeunes, qui souffrent moins d'effets secondaires, sont sollicités pour les essais, même si ces médicaments sont plus souvent destinés à des patients plus âgés ;

– les échelles de temps sont fréquemment manipulées, c'est-à-dire que les essais sont souvent de courte durée, même pour des médicaments administrés à vie ;

– comme ce sont les sociétés, et non les chercheurs, qui contrôlent l'analyse et la publication des données, les résultats des essais négatifs ou non concluants sont généralement dissimulés et des rapports sont rédigés pour présenter les produits sous un jour favorable.

Des incitations faussées

Sur le marché actuel de l'évaluation, les puissants tiennent les cordons de la bourse. Aucune évaluation n'est conçue et mise en œuvre sans la participation pleine et entière des responsables. Ces contraintes se traduisent par des incitations faussées qui menacent l'intégrité et l'indépendance de l'évaluation. Il n'est guère surprenant que les RCT soient favorisées par des intérêts particuliers, puisqu'elles se gardent bien d'examiner l'impact sur les résultats de l'aide d'une sélection inadéquate des programmes ou d'une mauvaise gestion.

Le cas de la recherche médicale démontre que les RCT sont exposées à une sélection trompeuse des comparateurs, à un tri sélectif des données, à des biais dans la présentation des résultats, à des pressions financières, etc. lorsqu'elles sont accaparées par des intérêts particuliers. Même si la recherche est menée par les universités, la plupart des essais sont désormais financés par des sociétés pharmaceutiques en vertu de contrats qui restreignent la liberté des chercheurs en permettant à des sponsors privés de contrôler étroitement les modèles d'évaluation, l'analyse des données, l'interprétation des recherches, la diffusion des résultats, etc.

Ainsi, la mainmise insidieuse des intérêts particuliers sur la recherche médicale démontre que les menaces qui pèsent sur la validité des évaluations peuvent être davantage liées à un manque d'indépendance qu'à une absence de rigueur méthodologique. En d'autres termes, la pratique de la recherche médicale n'a rien d'une norme d'excellence.

Compte tenu de l'influence croissante des intérêts commerciaux et géopolitiques dans le domaine de l'aide internationale, le triste bilan de la recherche médicale fait craindre des risques imminents pour l'évaluation du développement. Seuls des principes éthiques et des normes agréées permettant d'encadrer la pratique professionnelle peuvent faire obstacle à l'emprise des intérêts partisans sur l'évaluation.

Des contributions modestes aux connaissances sur le développement

En plus de vérifier si des interventions menées au niveau individuel « fonctionnent » comme prévu, les *randomistas* ont pour ambition de produire des résultats importants en matière de recherche en sciences sociales et de politiques. Selon le site web du J-PAL du MIT, « les évaluations randomisées peuvent générer des informations importantes sur le comportement humain et les institutions, en plus de mesurer les impacts de programmes et de politiques spécifiques. Les connaissances découlant de multiples évaluations randomisées sur un même sujet peuvent aider à la prise de décisions par des gouvernements, des ONG, des entreprises et des bailleurs de fonds qui travaillent à relever des défis similaires » (Dhaliwal et Olken, 2018). Tout porte à croire que ces prétentions ont une validité limitée.

Un champ d'application étroit

Les RCT considèrent leur absence de lien avec toute théorie comme un avantage. Ceci se transforme en inconvénient dans la recherche en sciences sociales, sauf si les RCT sont associées à d'autres méthodes et s'appuient sur des connaissances préalables (Vivalt, chap. 11, ce volume). En effet, les RCT individuelles ne peuvent pas à elles seules prétendre à la réplicabilité d'un contexte opérationnel à un autre. Les aléas statistiques associés à l'échantillonnage entravent sérieusement la transférabilité des résultats en dehors du contexte dans lequel les expérimentations ont été conçues et réalisées. Ce n'est pas seulement dû au fait que les RCT ne donnent pas toujours une estimation fiable des effets de traitement moyens, mais aussi parce que la garantie de causalité au niveau de l'intervention ne permet guère d'établir la validité externe des résultats des RCT (Deaton et Cartwright, 2018).

De surcroît, les RCT sont méthodologiquement parcimonieuses et ont une portée limitée. Comme elles s'attachent surtout à éliminer le biais de sélection des interventions de développement, elles ne traitent que de questions restreintes portant sur l'efficacité des mécanismes de fourniture de biens privés. Les biens publics, c'est-à-dire les biens qui ne sont ni rivaux ni exclusifs, ne se prêtent pas aisément à la randomisation.

Les RCT ne sont donc pas capables de traiter des enjeux cruciaux en matière de politique de développement, comme le changement climatique, la biodiversité, la sécurité publique, la propriété intellectuelle, etc. Pour ces biens, qui sont au cœur de la politique de développement durable, il n'est pas possible de concevoir des expérimentations permettant de faire la distinction entre ceux qui ont bénéficié du « traitement » et ceux qui n'en ont pas bénéficié.

Une position paternaliste

Les RCT se concentrent sur la manière dont les bénéficiaires de l'aide (c'est-à-dire les pauvres) pensent et se comportent. Cette position est en phase avec l'opinion selon laquelle la pauvreté est un choix personnel, plutôt que la conséquence des dispositifs sociaux et structures politiques existants. Les *randomistas* mènent un travail de terrain pour construire des enquêtes statistiquement plausibles. Mais ils privilégient des modèles mentaux préexistants et réactionnaires qui se focalisent sur des ajustements marginaux des politiques existantes plutôt que sur des choix radicaux de politiques alternatives.

Encouragés par les résultats de l'économie comportementale en vogue, ils sont ainsi enclins à remettre en question la rationalité des choix des pauvres et, plutôt que d'étudier les dysfonctionnements sociaux qui limitent leurs options et sapent leurs initiatives, ils se concentrent sur la façon dont les décideurs politiques peuvent les aiguiller vers des changements comportementaux prédéterminés, même si ces changements ne reflètent pas toujours leurs préférences ou leur situation.

Des contributions limitées à la connaissance

L'évaluation donne des résultats significatifs lorsqu'elle aborde des problématiques opérationnelles importantes et pertinentes. Pour porter ses fruits, elle doit donc passer par une sélection judicieuse des sujets d'évaluation. Sur le plan de l'utilisation, l'évaluation indépendante, conçue comme un outil d'apprentissage organisationnel et axée sur des questions stratégiquement pertinentes, présente des avantages majeurs par rapport aux évaluations expérimentales dispersées, réalisées dans des contextes très différents pour le compte de clients variés et souvent intéressés. C'est particulièrement le cas lorsque ces évaluations expérimentales sont mises en œuvre par des agents extérieurs, qui ont une expérience limitée du développement et qui sont handicapés par des asymétries informationnelles massives et poussés par des chercheurs impatients de publier leurs travaux.

Certes, les RCT contribuent à la connaissance du développement lorsqu'elles traitent d'une question politique pertinente, lorsqu'elles exploitent les résultats accumulés dans la littérature et lorsqu'elles sont complétées par des études observationnelles et des méthodes qualitatives. La Royal Swedish Academy of Sciences (2019), enthousiasmée par le savoir-faire expérimental de terrain affiché par les économistes du MIT et d'Harvard leur a ainsi attribué le prix Sveriges Riksbank 2019.

Les RCT ont par exemple contribué à infirmer les prétentions exagérées des ardents défenseurs du microcrédit, qui avaient vu dans les études de cas décrivant des programmes de microcrédit la clé de l'autonomisation des femmes et de la réduction de la pauvreté à grande échelle. Des RCT soigneusement construites dans divers contextes, associées à des observations sur le terrain, ont montré que le microcrédit est un produit financier certes utile, mais qu'il n'est nullement le garant d'un changement social radical.

Dans certains cas, les microprêts n'ont induit aucune différence notable dans l'influence des femmes sur les décisions et les dépenses des ménages. De même, les conditions rigides et les règles de prêt collectif destinées à protéger la viabilité financière des institutions de microcrédit se sont révélées peu adaptées aux besoins des entrepreneurs en herbe. Les programmes de formation commerciale mis en place par les microprêteurs pour aider les emprunteurs à développer leur entreprise n'ont pas non plus eu un impact significatif sur leurs bénéfices ou leurs ventes (Banerjee et Duflo, 2011). Les RCT ont ainsi contribué à démystifier certains des modèles en vogue, mais imparfaits, qui ont périodiquement déferlé dans le monde du développement.

Les RCT ont également « redécouvert » certaines bonnes pratiques bien établies dans le domaine du développement, notamment l'efficacité du rattrapage scolaire et des soins de santé préventifs, mise en exergue par la l'Académie royale des sciences de Suède. Dans la même veine, 58 RCT du Laboratoire d'action contre la pauvreté ont produit des données de terrain qui confirment les conclusions de praticiens expérimentés de la politique de l'éducation sur les facteurs permettant d'augmenter les inscriptions et la participation des étudiants, à savoir la baisse (ou la suppression) des frais de scolarité, la réduction des temps de trajet jusqu'à l'école, la prise en charge des problèmes de santé des enfants et l'information des parents sur les bénéfices de l'enseignement.

Dans le même ordre d'idées, une expérimentation de terrain menée dans 100 villages indiens a permis de valider les résultats d'études antérieures sur le développement agricole : les journées passées au champ par les agriculteurs sont utiles et rentables pour la diffusion de nouvelles variétés à haut rendement. En outre, et sans surprise, une étude expérimentale complexe réalisée au Kenya a confirmé que les conseils sur les applications d'engrais devaient être guidés par la maximisation des profits au niveau de l'exploitation agricole, plutôt que par la maximisation des rendements. C'est comme si les *randomistas* cherchaient des preuves que les sciences économiques sont dignes d'intérêt ou que leur instrument d'évaluation favori « fonctionne ».

Les RCT ne sont qu'un outil parmi d'autres

Compte tenu de ces observations, les RCT obéissent-elles aux principes, objectifs et pratiques de base de la discipline de l'évaluation ? Si les définitions de l'évaluation et des modèles d'évaluation sont légion, nombreux sont ceux qui reconnaissent le rôle essentiel de la *valeur* dans l'évaluation, au sens proposé dans la définition concise donnée par SCRIVEN (1991 : 5) et largement acceptée par la communauté de l'évaluation comme « le processus de détermination du mérite, de la pertinence et de la valeur des choses – ou le résultat de ce processus ». Les trois dimensions d'intérêt de cette définition sont interdépendantes, mais c'est le critère de valeur qui distingue le plus l'évaluation des autres types d'enquête.

Tout d'abord, le *mérite* détermine les performances par rapport aux normes de qualité. Il s'agit *de bien faire les choses* pour atteindre les objectifs de l'intervention, ce qui relève de l'*efficacité* définie dans le glossaire du Comité d'aide au développement (2010 : 20) comme étant la « mesure selon laquelle les objectifs de l'action de développement ont été atteints, ou sont en train de l'être, compte tenu de leur importance relative ».

Ensuite, l'*intérêt* a trait au fait de *faire les choses appropriées*. Il fait référence aux bénéfices nets qui peuvent légitimement être attribués à l'intervention en tenant compte de considérations de mérite fondées sur les perspectives de ceux qui sont censés bénéficier de l'intervention et d'autres parties prenantes, personnes ou entités concernées par l'intervention. Il s'agit de la *pertinence* telle que définie dans le glossaire du comité d'aide au développement comme étant « la mesure selon laquelle les objectifs de l'action de développement correspondent aux attentes des bénéficiaires, aux besoins du pays, aux priorités globales, aux politiques des partenaires et des bailleurs de fonds » (Comité d'aide au développement, 2010 : 32).

Enfin, la *valeur* évoque l'intérêt collectif et intègre également des considérations d'économie dans les ressources utilisées pour atteindre les résultats escomptés, c'est-à-dire faire les choses efficacement par rapport à d'autres façons de concevoir et de mettre en œuvre l'intervention. Plus précisément, l'*efficience* est définie par le glossaire du comité d'aide au développement comme étant « la mesure selon laquelle les ressources (fonds, expertise, temps, etc.) sont converties en résultats de façon économe » (Comité d'aide au développement, 2010 : 21).

Dans quelle mesure les RCT sont-elles évaluatives ?

Les RCT font partie intégrante de la boîte à outils de l'évaluateur et il ne fait guère de doute que la détermination de la causalité des résultats observés (l'objectif fondamental des RCT) constitue un élément essentiel pour juger de leur mérite. Par contre, cette approche restreinte de l'évaluation ne permet guère de définir si une intervention est pertinente, efficiente ou durable. Établir qu'une intervention fonctionne n'est pas la même chose que déterminer si c'était la bonne intervention, comprendre pourquoi elle a fonctionné de cette façon ou si ses objectifs valaient la peine d'être poursuivis au départ.

Les objectifs, la portée, la structure et le contexte du programme sont d'une grande importance pour façonner le résultat des politiques et programmes. Même lorsque les expérimentations constituent une approche adaptée pour l'analyse d'attribution, les résultats peuvent ne pas satisfaire tous les besoins ressentis par les décideurs politiques, qui se préoccupent moins de ce qui s'est passé dans l'expérimentation que des chances qu'elle a de continuer à fonctionner dans d'autres contextes ou dans le futur, étant donné la prédominance des environnements de mise en œuvre complexes et volatils (Cartwright et Munro, 2010).

Enfin, en l'absence de théorie susceptible d'être infirmée, il n'est pas possible de faire progresser les connaissances. Une évaluation de grande qualité, permettant d'établir la validité de la théorie sur laquelle repose le programme, requiert une compréhension approfondie du fonctionnement de celui-ci. La bonne compréhension des relations causales et l'identification des explications concurrentes à réfuter nécessitent de bien appréhender l'intervention, sa conception, ses protocoles de mise en œuvre et les motivations des participants et des bénéficiaires du programme. Les questions ouvertes et les approches qualitatives sont mieux adaptées à ces questions.

Cela explique pourquoi l'évaluation indépendante, fondée sur le travail de terrain, intégrée à l'organisation et réalisée par des praticiens expérimentés, s'est avérée beaucoup plus efficace que les RCT pour réorienter les processus opérationnels et pour supprimer (Gautam, 2000) ou restructurer les lignes de crédit de développement inefficaces (Tendler, 1993). La caricature de l'évaluation interne inévitablement subordonnée à l'intérêt personnel des institutions n'est pas davantage valable, surtout lorsque la fonction d'évaluation est responsable devant l'autorité suprême de l'organisation plutôt que devant la direction opérationnelle et qu'elle est mandatée pour attester de la qualité de processus auto-évaluatifs (Picciotto, 2013).

Les politiciens et les fonctionnaires font des choix collectifs concernant l'allocation et l'utilisation des ressources publiques. Ils ont pour mission de valoriser au mieux l'ensemble des actifs dont ils ont la charge. Ils doivent démontrer qu'ils le font de manière responsable et efficace. Par conséquent, la clé pour légitimer le pouvoir et l'autorité tient à un argumentaire valide et digne de foi sur la création de valeur publique.

Dans sa mission *sommative*, l'évaluation examine les résultats des politiques et des programmes, et s'attache à déterminer dans quelle mesure les autorités dirigeantes ont agi de manière responsable. Le principal mécanisme permettant de remédier aux mauvaises performances d'un gouvernement est la voix des citoyens. L'évaluation l'amplifie en fournissant aux électeurs des connaissances pertinentes sur les performances du secteur public.

La gestion du secteur public a longtemps été dominée par une évaluation de la valeur publique basée sur de simples mesures de rendement et des coefficients budgétaires plutôt que sur les résultats et les impacts. Or, ces indicateurs laissent beaucoup à désirer. Ils ne mesurent pas les résultats et sont faciles à manipuler. Les informations fournies par les responsables du secteur public sur leur travail

nécessitent une validation étayée : l'évaluation indépendante dans le secteur public équivaut à la vérification des comptes dans le secteur privé.

C'est donc là qu'intervient l'évaluation indépendante *:* elle a pour mission de déterminer de manière fiable si les erreurs décisionnelles sont dues à des circonstances sur lesquelles les décideurs n'ont aucune prise ou si les risques encourus pouvaient être mieux gérés. Une évaluation juste et objective est un facteur de responsabilisation : elle garantit que les promesses faites par les politiciens et les décideurs des secteurs public, privé et associatif sont systématiquement mises en regard des résultats fournis par des processus d'évaluation justes et objectifs. La mise en regard des résultats avec les promesses faites lors du lancement d'une politique ou d'un programme fait partie intégrante du processus démocratique.

Les méthodes orientées sur les objectifs occupent ainsi une place privilégiée dans l'arsenal de l'évaluateur. Mais, à cet égard, l'évaluation expérimentale ne peut prétendre faire la distinction entre les effets imputables aux différents acteurs – qui sont invariablement impliqués dans les interventions des politiques – et aux programmes. Pourtant, la plupart des politiques et des programmes sociaux reposent sur des *partenariats* entre diverses entités du gouvernement, du secteur privé et de la société civile pour atteindre des résultats et avoir des impacts. Si l'on n'évalue pas les contributions respectives des partenaires et le respect de leurs obligations réciproques, les responsabilités de chacun restent floues.

La responsabilité d'un échec peut par exemple être totalement éludée si elle est sommairement attribuée à la mauvaise performance d'un partenaire. Inversement, la responsabilité d'une réussite peut être injustement attribuée à un seul partenaire (une agence gouvernementale, par exemple), que sa contribution à l'atteinte des objectifs communs le justifie ou non. L'absence d'évaluation adéquate peut donc avoir des effets délétères sur les motivations en émettant des signaux erronés.

Ainsi, lorsque l'échec d'un programme ou d'un projet (s'il se produit) est intégralement imputé à l'organisme qui met en œuvre l'intervention (indépendamment des influences exogènes et des contributions des partenaires), cela induit une aversion au risque et peut même conduire à la suspension des programmes qui ne parviennent pas à atteindre des objectifs ambitieux, manquant par là même l'opportunité de les adapter pour qu'ils puissent réussir.

Les bonnes évaluations doivent donc explicitement tenir compte des responsabilités respectives et des obligations réciproques des partenaires. Si les performances des différents acteurs ne sont pas évaluées séparément pour expliquer les résultats et les impacts, le risque moral l'emportera. Une évaluation sommative de grande qualité ne se contente donc pas de répondre à la question de savoir si une politique ou un programme fonctionne ou non, ce qui est l'objectif très limité de l'évaluation d'impact expérimentale.

En résumé, pour une approche se targuant souvent à tort de favoriser fortement la responsabilisation, la nouvelle conception de l'évaluation d'impact utilisant

les RCT élude la question embarrassante de savoir qui doit répondre des écarts observés entre, d'un côté, les objectifs des politiques et des programmes et, de l'autre, les résultats réels. En se limitant à l'*attribution* d'effets à l'intervention, les RCT ne répondent pas à la question de la *contribution*, c'est-à-dire dans quelle mesure chacun des partenaires individuels impliqués dans l'action de développement a contribué à la réalisation des objectifs du programme ou du projet, et ce qui pourrait être fait pour améliorer ses performances.

Manier les bons outils

Les RCT ne sont qu'un outil d'évaluation parmi d'autres. En tant que telles, elles ne devraient pas être autorisées à dominer ce qui est avant tout un processus créatif, analytique et participatif. Les méthodes expérimentales présentent de nombreuses caractéristiques statistiques que d'autres modèles d'évaluation peuvent difficilement égaler dans certaines circonstances. Il est dangereux pour la bonne gestion de l'évaluation de surinvestir une technique unique. Un outil ne peut remplir que la ou les fonction(s) pour laquelle/lesquelles il a été conçu.

Utiliser les bons outils, et les utiliser avec soin et compétence, compte pour beaucoup dans la qualité de l'évaluation. Mettre en œuvre des méthodes inappropriées peut ruiner une évaluation. Mais d'autres facteurs peuvent également compromettre la rigueur d'une évaluation : une collecte de données bâclée, des évaluations politiquement naïves, le manque d'indépendance, l'absence de compétences adéquates des évaluateurs, l'incapacité à se focaliser sur l'utilisation, l'ignorance du contexte, une participation limitée des parties prenantes, la concentration sur des questions sans importance ou non pertinentes, etc.

Des outils bien choisis et employés dans les règles de l'art contribuent à la validité des évaluations. Ils facilitent la comparaison de celles-ci, ainsi que leur chiffrage financier et leur planification. Ils rendent les résultats des évaluations plus crédibles et plus prévisibles. Comprendre et mesurer les limites des outils utilisés dans un contexte est essentiel pour garantir la qualité. L'incapacité à corréler le modèle détaillé de l'évaluation avec les questions prioritaires identifiées au stade de la planification explique pourquoi de nombreuses évaluations tournent mal.

La compréhension des forces, faiblesses et limites respectives des méthodes et outils d'évaluation constitue par conséquent une compétence critique pour les évaluateurs. Si les méthodes expérimentales et quasi expérimentales peuvent, dans certaines circonstances, éclairer l'attribution des résultats observés, les études observationnelles basées sur la théorie et les évaluations de processus appliquant une triangulation judicieuse des méthodes sont plus adaptées pour expliquer le « comment » et le « pourquoi » des effets observés. C'est donc une chance que toutes les directives et normes d'évaluation nationales et régionales accordent aux approches qualitatives l'importance et la crédibilité qui se doit. Elles privilégient ainsi la pertinence méthodologique et le pluralisme par rapport à l'orthodoxie doctrinale.

Conclusion

Le mouvement expérimentaliste a des racines historiques profondes. Vendues avec succès par des chercheurs-entrepreneurs, les RCT jouissent d'une grande loyauté de leurs praticiens. Elles promettent certitude et rigueur dans un monde du développement caractérisé par une volatilité et une complexité extraordinaires. Elles comportent pourtant une foule de limites. Elles sont coûteuses et doivent faire face à de nombreux défis statistiques et éthiques. Leurs fondements épistémologiques sont peu solides, leur prétention au titre d'étalon-or est injustifiée et les arguments selon lesquels les procédures d'évaluation randomisée qui ont fait leurs preuves dans le secteur de la santé sont la clé de la rigueur de l'évaluation dans le domaine des sciences sociales sont infondés.

Pour ce qui est des interventions menées au niveau individuel, les RCT permettent de tirer des conclusions d'attribution seulement dans le cas d'interventions simples et mises en œuvre dans des environnements stables. Elles ne contribuent à la recherche sur l'action publique en général que lorsqu'elles s'inscrivent dans un processus de production de connaissances cumulées qui s'appuie également sur des études observationnelles et des études qualitatives. D'autres méthodes d'évaluation, associées ou non à l'expérimentation, permettent de traiter de façon convaincante les questions complexes d'une activité de développement qui ne cesse d'évoluer.

En tant qu'évaluations, les RCT ne traitent qu'un seul des critères évaluatifs fondamentaux (l'efficacité), que les interventions des politiques et des programmes doivent satisfaire pour être jugées efficaces. Elles échouent à traiter les questions de pertinence, d'efficience et de durabilité, qui sont souvent plus importantes. Elles ne permettent pas non plus de distinguer et d'estimer les contributions de chaque partenaire responsable de la réussite ou de l'échec de ces interventions, une lacune majeure puisque la responsabilité envers les citoyens fait partie intégrante de la mission d'évaluation.

Les *randomistas* ne sont donc pas des évaluateurs, puisque les RCT ne sont pas des évaluations. Néanmoins, les RCT continueront à jouer un rôle majeur dans le domaine du développement, puisqu'elles sont solidement ancrées dans le monde de la recherche, qu'elles apportent de modestes contributions à la connaissance du développement, qu'elles ne remettent pas en cause les prérogatives du pouvoir. En dépit de leur portée limitée, elles répondent à une réelle demande de preuves officiellement plausibles quant au « bon fonctionnement » des interventions de développement. Le prix Nobel obtenu en 2019 renforcera davantage encore le rôle privilégié des études expérimentales dans l'économie du développement.

Remerciements

Lant Pritchett a formulé des commentaires judicieux sur une version antérieure de ce chapitre, mais il n'est nullement responsable de ses erreurs et omissions.

Partie 4

Quelques pistes de réflexion (ciblées) : éthique et méthode

Chapitre 10

Expérimentations aléatoires et éthique

Les économistes doivent-ils se soucier de l'équipoise ?

Michel Abramowicz et Ariane Szafarz

Introduction

Lors d'évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT), est-il admissible de proposer, pour des raisons financières, des options inégalement porteuses au groupe traité et au groupe contrôle ? Tandis que les médecins répondent négativement à cette question, certains économistes semblent tentés de dire « oui ». Même si l'équipoise constitue une composante majeure des RCT médicales, il est frappant de constater que de nombreux économistes n'en ont jamais entendu parler. Ce chapitre entend combler cette lacune en examinant comment la littérature scientifique médicale formalise le principe d'équipoise et, de là, déterminer comment ce principe pourrait être adapté aux RCT dans le domaine de l'économie.

L'équipoise est définie par Freedman (1987 : 141) comme un « état de réelle incertitude de la part du chercheur clinicien concernant les mérites thérapeutiques relatifs de chaque branche de l'alternative ». L'auteur considère ce principe comme « un prérequis éthique indispensable à toute recherche clinique ». L'équipoise exige qu'avant le début de l'expérimentation, il y ait, pour chacun des traitements étudiés, un degré d'ignorance identique des avantages et inconvénients. Cette exigence naît de l'injonction éthique selon laquelle, en cas de supériorité de l'un des traitements, on lèse les patients qui recevraient l'autre option. Et, puisque les expériences médicales sont réalisées en double aveugle, le non-respect de l'équipoise peut nuire à tous les sujets de l'étude. En ce sens, l'exigence d'équipoise renforce la déclaration d'Helsinki de 1964

de l'Association médicale mondiale[1], qui stipule notamment que les groupes contrôle doivent recevoir le meilleur traitement disponible. Cette exigence est absolue, elle doit être appliquée quelles que soient les conditions spécifiques de l'étude, y compris sa localisation géographique.

Pourtant, l'équipoise reste un sujet controversé, car sa mise en œuvre pratique soulève des questions essentielles concernant, en particulier, le délicat équilibre à trouver entre les opinions et préférences des cliniciens, des chercheurs et du patient traité (Lilford et Jackson, 1995). De toute évidence, l'appréciation des mérites thérapeutiques peut varier en fonction de la sensibilité des parties prenantes, et ainsi conduire à des dilemmes éthiques (Schafer, 1982), tels que l'épineuse mise en balance des devoirs du médecin envers son patient et des progrès de la science (Botros, 1990). Miller et Joffe (2011) contestent toutefois que les enjeux soient limités à la relation médecin-patient. Les auteurs placent le débat dans le contexte plus large de la politique de santé. Ils mettent en parallèle l'intérêt des patients et la nécessité de connaissances pour l'approbation de médicaments. Ce faisant, ils lient l'équipoise à l'arbitrage, majeur en santé publique, qui oppose la liberté individuelle à la justice sociale (Kass, 2001 ; Childress *et al.*, 2002). À cet égard au moins, le lien est évident entre l'éthique des RCT dans les domaines de la médecine et de l'économie.

Alors que la littérature médicale débat âprement de la pertinence des diverses spécifications du principe d'équipoise, la recherche économique reste muette sur le sujet. Bien entendu, les RCT en économie sont examinées par des comités d'éthique. Mais ces comités sont généralement locaux. Il n'y a toujours pas d'exigences éthiques à grande échelle, et encore moins de références à l'équipoise. Notre objectif est de briser l'apparente indifférence des économistes envers une question éthique, pourtant essentielle pour l'expérimentation médicale. Dans la lignée de Baele (2013) et Petticrew *et al.* (2013), qui préconisent le développement d'une « équipoise sociale », ce chapitre entend lancer un débat sur l'équipoise au sein de la communauté des économistes qui utilisent les RCT.

Qu'est-ce que l'équipoise ?

La mobilisation d'êtres humains comme sujets d'expériences crée des problèmes éthiques délicats. La pratique consistant à tester les traitements médicaux à l'aide d'expériences contrôlées remonte à la nuit des temps, mais le principe d'équipoise est nettement plus récent. Il a été conceptualisé au XX^e^ siècle avec l'émergence de la randomisation avec groupes contrôle et attribution à l'aveugle du traitement ou du placebo (Di Tillio *et al.*, 2017).

1. Carlson *et al.* (2004) présentent les révisions ultérieures de cette déclaration. Voir également les Directives éthiques internationales de 1982 du Council for International Organizations of Medical Sciences pour la recherche biomédicale impliquant des sujets humains (CIOMS, 2002).

Les codes modernes d'éthique médicale sont généralement guidés par les principes du serment d'Hippocrate énonçant les devoirs du médecin envers son patient (Orr *et al.*, 1997 ; Miles, 2005), comme l'obligation de dispenser le meilleur traitement possible. Ainsi, si le médecin a de bonnes raisons[2] de croire que le traitement A est meilleur que le traitement B, alors il ne peut prescrire B au lieu de A à aucun de ses patients (Shaw et Chalmers, 1970). De même, il doit s'abstenir de participer à toute étude scientifique qui le conduirait à administrer le traitement B plutôt que le traitement A. Cette restriction majeure, imposée par l'éthique médicale, peut entraver le développement d'études médicales à grande échelle comparant les traitements A et B. C'est pour aborder ce problème que Freedman (1987) introduisit le concept d'« équipoise clinique », qui impose de disposer de preuves statistiques suffisantes pour pouvoir conclure que le traitement A ne domine pas le traitement B[3]. L'idée sous-jacente est de placer l'évaluation des mérites thérapeutiques respectifs des deux traitements sous la responsabilité d'un groupe d'experts médicaux, plutôt que sous celle d'un seul individu (Freedman, 1987).

Sans compromettre les fondements des RCT, l'équipoise clinique a permis d'assouplir les contraintes pratiques qui avaient parfois dicté l'arrêt prématuré d'études dont les premiers résultats avaient indiqué qu'un traitement était meilleur que l'autre, du moins à court terme. L'idée de Freedman était de laisser suffisamment de temps à la communauté scientifique pour consolider les résultats de vastes études. Freedman défend que l'équipoise clinique permet aussi de réduire le problème de la faible participation aux études, due à la réticence des médecins à enrôler leurs patients dans des études qui les gênent (Taylor *et al.*, 1984).

Dans l'ensemble, le principe opérationnel de l'équipoise clinique s'est révélé fructueux pour le développement d'études médicales à grande échelle, en contribuant tant à la conception qu'à la faisabilité des RCT. Ce faisant, l'équipoise permet de concilier les droits des participants aux études et la quête du progrès scientifique (London, 2017).

Les premières étapes de la mise en œuvre du principe d'équipoise en recherche médicale aident les économistes à saisir les défis et enjeux des RCT. Ainsi, la communauté des chercheurs et cliniciens en cardiologie fut la première à lancer des RCT médicales à grande échelle et à visées thérapeutiques. L'une de ces études pionnières, la *Beta-Blocker Heart Attack Trial* (BHAT, 1982) concerne les bêta-bloquants, qui sont des médicaments mis au point dans les années 1960 et dont la découverte a été couronnée en 1988 d'un prix Nobel de médecine

2. L'avis du médecin se fonde sur son expérience et son analyse personnelle de la situation. La composante subjective complique inévitablement la formalisation du principe d'équipoise.

3. Les économistes du développement pourraient arguer que certaines populations pauvres n'ont accès ni au traitement A, ni au traitement B, de sorte que même B améliorerait leur situation. La section « Les économistes doivent-ils se soucier de l'équipoise ? » pondère cet argument souvent utilisé pour justifier l'attribution de traitements moins favorables à certains groupes de contrôle. Nous soutenons que comparer les conditions de vie sous une RCT à celles de la vie normale conduit à ignorer le rôle des investigateurs, qui peuvent influencer notablement les comportements et le ressenti des participants. Sur ce point, les RCT en sciences sociales doivent, comme les RCT médicales, bannir toute relativisation des exigences éthiques en fonction des populations étudiées.

décerné à Sir James Black. Ces médicaments étaient initialement prévus pour traiter l'hypertension. L'étude BHAT a randomisé des sujets ayant survécu à un récent infarctus du myocarde en leur administrant soit du propranolol, le premier bêta-bloquant largement disponible sur le marché, soit un placebo (YUSUF *et al.*, 1985). Les résultats ont montré une forte réduction (7,2 % contre 9,8 %) de la mortalité totale à moyen terme (le suivi moyen était de 24 mois). À ce jour, trente-six ans après la publication de cette étude fondamentale, les bêta-bloquants constituent toujours un élément primordial de la prévention secondaire après un infarctus du myocarde. Les bêta-bloquants sont également les médicaments les plus actifs contre l'angine de poitrine – une affection chronique douloureuse causée par des artères cardiaques partiellement obstruées qui, une fois totalement bouchées, entraînent généralement un infarctus – et ils réduisent considérablement la mortalité due à l'insuffisance cardiaque (MCMURRAY, 2010).

Avant la mise sur pied de l'étude BHAT, il y avait une réelle équipoise au sein de la communauté médicale clinique et scientifique sur la protection potentielle offerte par le bêta-blocage après un infarctus du myocarde (NIES *et al.*, 1973 ; SHAND, 1975). Même si les expériences animales avaient montré une amélioration de la survie et si les patients souffrant d'angine de poitrine supportaient bien le médicament, on redoutait que la baisse de la pression artérielle associée aux bêta-bloquants soit néfaste (THEROUX *et al.*, 1974).

Le degré d'incertitude relatif à la supériorité d'un traitement peut cependant varier au cours d'une RCT pour plusieurs motifs, dont les résultats partiels de l'étude elle-même et la publication de nouveaux résultats par d'autres équipes. Les structures chargées de la supervision des études (*Data and Safety Monitoring Boards* – DSMB) ont la responsabilité de déterminer si l'équipoise a été suffisamment perturbée pour justifier l'arrêt de l'étude (DICKERT et EMANUEL, 2015 : 31). C'est une décision difficile, car, d'une part, un arrêt prématuré peut nuire à la validité globale de l'étude et, d'autre part, la poursuite de l'étude avec une équipoise compromise peut nuire à ses sujets. Par exemple, des rumeurs ont accusé l'étude ISIS-2 (1988) d'avoir ignoré les résultats intermédiaires, publiés en 1987, d'une équipe italienne concurrente, le Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). L'étude ISIS-2 portait sur 17 187 patients admis dans une unité coronaire avec un diagnostic d'infarctus aigu du myocarde. Les patients ISIS-2 ont été randomisés et se sont vu administrer soit de la streptokinase, un médicament qui dissout les caillots et dont on espérait qu'il améliore le pronostic en réduisant la taille de l'infarctus, soit un placebo. Au cours de l'étude ISIS-2, des résultats partiels du GISSI favorisant fortement la streptokinase ont commencé à poindre, suggérant qu'ISIS-2 lésait ses patients sous placebo, c'est-à-dire le groupe contrôle, par rapport à ceux sous traitement. Cet épisode s'est déroulé entre 1985 et 1988, une période durant laquelle le concept d'équipoise était encore peu connu dans le milieu de la recherche médicale. Finalement, l'étude ISIS-2 s'est poursuivie comme prévu et ses résultats ont confirmé les avantages substantiels de la streptokinase en termes de mortalité. Aujourd'hui, trente ans après ses débuts, l'équipoise fait partie intégrante des normes éthiques de base pour les RCT médicales.

Équipoise *versus* double aveugle

Même si la communauté médicale unanime place l'équipoise parmi ses normes éthiques, la mise en œuvre de ce principe soulève plusieurs questions pratiques. Le problème le plus basique, et peut-être le plus délicat, est de déterminer comment on peut prouver qu'une étude donnée satisfait l'équipoise clinique. Plusieurs méthodes peuvent être utilisées comme, par exemple, se référer à des études antérieures ou mettre en exergue les divergences de vues au sein de la communauté clinique concernant les traitements existants. Certaines circonstances peuvent toutefois compromettre la mise en œuvre de l'équipoise clinique. Cette section s'intéresse à deux préoccupations majeures, qui peuvent sembler familières aux chercheurs en sciences sociales : l'absence d'aveugle expérimental et la décision d'enrôler un patient souffrant d'une pathologie préexistante.

Le double aveugle est devenu la norme dans les RCT médicales. En revanche, il n'est guère pratiqué dans les RCT économiques pour d'hypothétiques raisons techniques. Cette section montre que, comme en économie, la mise en œuvre d'expériences en aveugle n'est pas toujours faisable dans certaines branches de la médecine, comme la chirurgie[4]. Mais, suivant en cela la littérature médicale, nous soutenons que, même sans aveugle, le principe d'équipoise peut être appliqué. Par exemple, l'étude GISSI (1987) était une RCT sans aveugle « dont le protocole spécifiait trois analyses intermédiaires, avec 3 000, 6 000 et 9 000 patients recrutés. Les résultats de ces analyses n'ont été présentés qu'au comité d'éthique ; une différence de mortalité supérieure à trois écarts-types ou une incidence anormalement élevée de réactions indésirables à la streptokinase aurait conduit le comité à ordonner l'arrêt anticipé de l'étude » (GISSI, 1987 : 398). Cette règle d'arrêt montre que l'absence d'aveugle expérimental fait de l'équipoise une exigence encore plus impérative.

De nos jours, les RCT médicales sont généralement conduites en aveugle. Dans une étude en simple aveugle, le patient ne sait pas quel traitement lui est administré. Le double aveugle signifie que le sujet et le clinicien de terrain (qui traite les patients, les surveille et transmet les résultats au comité directeur de l'étude) ignorent tous deux qui reçoit quel traitement (DAYS et ALTMAN, 2000), le but étant de limiter les biais. Pour le clinicien, le fait de savoir quel traitement est appliqué peut inconsciemment altérer son interprétation des résultats. Quant au patient, son expérience subjective peut être influencée non seulement par sa connaissance du médicament administré, mais aussi par son ressenti des attentes du médecin. Les patients ont souvent tendance à penser que le médicament expérimental est plus efficace que le médicament de référence ou le placebo – même si, avec l'équipoise, ils se trompent – et donc à déclarer moins de symptômes lorsqu'ils pensent bénéficier de l'option perçue comme meilleure. PODSAKOFF *et al.* (2003) résument les sources potentielles de biais couramment observés dans la recherche

4. YOUNG *et al.* (2004) suggèrent d'apprécier la faisabilité d'une RCT avec équipoise en chirurgie en organisant des sondages parmi les chirurgiens du domaine. Ces sondages viseraient à évaluer la possibilité d'opposer deux traitements chirurgicaux.

comportementale. Ainsi, le biais de la désirabilité sociale procède du désir des sujets d'une étude de réagir selon l'acceptabilité sociale perçue plutôt que selon leurs sentiments. Les mêmes biais s'appliquent, dans une certaine mesure, aux cliniciens. On notera que le simple aveugle peut favoriser – souvent de manière inconsciente – les biais d'autocomplaisance des chercheurs (CAMFIELD *et al.*, 2014), d'où l'avantage du double aveugle. Cet argument s'applique tout particulièrement aux études dont les résultats sont intangibles ou subjectifs, qui sont monnaie courante en sciences sociales. Plus généralement, plus le résultat final est précis, moins l'aveugle est nécessaire. Pour donner un exemple extrême, il est difficile de mal interpréter la mort d'un patient. Mais dans tous les cas, lorsqu'il est faisable, le double aveugle est à présent la norme dans les RCT médicales.

Si la plupart des RCT médicales portent sur des médicaments, certaines évaluent des interventions sur le mode de vie ou des procédures chirurgicales. Pour ces études, trouver un placebo compatible avec l'utilisation du double aveugle, voire du simple aveugle, est loin d'être facile. Les études basées sur la chirurgie soulèvent la question délicate de pratiquer ou pas une opération fictive. Si les patients du groupe contrôle ne sont pas opérés, ils sauront immédiatement qu'ils ne sont pas traités et l'effet placebo est perdu. Une solution alternative consiste à amener tous les patients en salle d'opération, de sorte que les patients non opérés qui entrent en salle de réveil sont encore à moitié anesthésiés et ont une nouvelle cicatrice. Le second scénario renforce indéniablement la validité de l'étude, puisque la seule différence entre le groupe contrôle et le groupe traité est le traitement chirurgical réalisé. Néanmoins, cette méthode porte un préjudice bien réel aux sujets non traités en les soumettant à des risques chirurgicaux (bien que non thérapeutiques) et en les anesthésiant inutilement. Ce dilemme traduit l'arbitrage entre l'intérêt individuel et le bien général – l'équipoise *versus* le simple aveugle. En sciences sociales, où l'aveugle expérimental est difficile, voire impossible, à mettre en pratique, la balance devrait, paradoxalement, pencher du côté de l'équipoise.

Au-delà de l'argument purement éthique, l'existence de l'effet dit « Hawthorne » montre que l'absence d'aveugle expérimental renforce la nécessité de faire appel au principe d'équipoise. Cet effet, défini par les sociologues ROETHLISBERGER et DICKSON (1939), fait référence à la situation observée dans les usines Hawthorne de la Western Electrical Company à Chicago entre 1924 et 1927. En testant l'impact de l'intensité de l'éclairage artificiel de l'usine sur la productivité des travailleurs, les chercheurs ont obtenu un effet positif tant suite à une augmentation que suite à une réduction d'intensité. L'impact psychologique du changement des conditions de travail est donc apparu plus important que la nature même du changement, puisque l'impact ne peut affecter que le groupe traité où les personnes observent le changement de leurs conditions de travail, par opposition au groupe contrôle où rien ne change. Par conséquent, ne pas offrir de placebo aux individus du groupe contrôle peut non seulement leur porter préjudice, mais aussi nuire à la fiabilité de l'étude en introduisant des faux positifs dus à l'effet Hawthorne. Un nombre croissant d'économistes rend conscience de cette problématique et soumet dès lors le groupe contrôle à quelque changement.

L'inclusion éventuelle, dans une étude, d'un patient atteint d'une affection préexistante amplifie la tension morale qui oppose le devoir du médecin envers son patient à l'objectif de progrès scientifique. Selon WEIJER *et al.*, (2000 : 756), la solution retenue dépendra du côté de l'Atlantique où l'on se situe. « Au Royaume-Uni, le principe de l'incertitude individuelle est largement admis. Cependant, en Amérique du Nord, l'équipoise clinique – reflétant l'incertitude collective – constitue la base éthique dominante. » Le principe dit d'incertitude individuelle fonctionne comme une équipoise qui tient compte de l'état de santé préexistant du patient. En vertu de ce principe, ce sont les intérêts individuels de chaque patient que le médecin examinera avant d'envisager sa participation à l'étude. Ce principe peut compliquer le recrutement de participants, ce qui risque alors de compromettre l'équipoise clinique de l'étude.

En revanche, la Commission consultative nationale de bioéthique des États-Unis (National Bioethics Advisory Commission – NBAC) impose aux comités institutionnels d'éthique (CIE) « de déterminer si la relation entre les risques et avantages potentiels est raisonnable. Pour ce faire, les CIE doivent déterminer si les procédures répondent aux critères d'équipoise scientifique […] en plus d'être justifiées par l'accroissement des connaissances pour la société. Les chercheurs et les CIE doivent comprendre que le terme d'*équipoise scientifique* s'applique à tout type de recherche impliquant des interventions ou des procédures qui offrent la perspective d'un bénéfice direct pour les participants […] » (NBAC, 2001 : 77). Les sujets dont l'affection préexistante possède un traitement reconnu tireront profit d'études qui opposent le traitement reconnu davantage comme un placebo, d'une part, au traitement reconnu comme le médicament expérimental, d'autre part.

Les économistes doivent-ils se soucier de l'équipoise ?

Alors que les économistes organisent volontiers des RCT où les membres du groupe contrôle ne reçoivent rien, les médecins considèrent qu'un placebo – qui est déjà mieux que rien – ne constitue pas une option éthiquement admissible lorsque l'état le plus récent de la science offre un meilleur traitement. Ce principe d'équipoise s'applique même si le meilleur traitement est onéreux et que les sujets testés ne pourraient pas se le payer.

Comment se fait-il que la question éthique relative à l'application de l'équipoise soit restée si marginale chez les économistes qui organisent des RCT ? Au vu de l'émergence relativement récente de ces RCT, il se pourrait que la recherche économique n'en soit encore qu'au stade pré-Freedman. Cette hypothèse n'est cependant guère crédible, puisque les économistes se sont abondamment inspirés des expériences médicales pour concevoir les leurs.

Les chercheurs des communautés tant médicales qu'économiques connaissent donc vraisemblablement les controverses relatives aux études médicales non éthiques (HALPERN *et al.*, 2002), y compris celles organisées dans les pays en développement (GULHATI, 2004 ; JINTARKANON *et al.*, 2005 ; MILFORD *et al.*, 2006)[5]. Et pourtant, les économistes et les chercheurs en sciences sociales ont tendance à ignorer l'exigence d'équipoise en désavantageant généralement le groupe contrôle.

Voici, à titre d'exemple, le récent débat ayant opposé la professeure Megan Stevenson, de l'université George Mason, qui étudie l'impact des cautions pénales à l'aide de RCT, au Massachusetts Bail Fund (MBF), un fonds qui verse jusqu'à 500 dollars de caution en faveur de personnes à faible revenu. L'extrait suivant de leur conversation sur Twitter incarne l'essence même de la controverse sur l'équipoise[6]. On observe en effet que chacune des parties y développe les arguments clés du débat.

> « Les RCT attribuent de manière aléatoire un "nouveau traitement" (dans ce cas-ci, l'aide d'un fonds de cautionnement) à un groupe donné, tandis qu'un autre groupe reçoit le "traitement standard" (dans ce cas-ci, aucune aide pour la caution, ce qui entraîne l'incarcération préventive). Vous êtes déjà horrifiés ? NOUS CONNAISSONS L'IMPACT DE LA CAUTION PÉNALE. La recherche EXISTE : les gens ont de meilleures chances et de meilleures perspectives dans la vie lorsqu'ils peuvent défendre leur cause en étant libres. Nous croyons les gens quand ils nous disent la différence que cela fait pour eux de voir leur caution payée. C'est une FAUTE de désigner au hasard ceux qui reçoivent un traitement qui leur sauvera la vie tout en en envoyant d'autres en prison. Quand on connaît les disparités raciales profondes et flagrantes de nos tribunaux et nos prisons, faire ce type de "recherche" relève du RACISME » (MBF, 8 mars 2019).
>
> « Tant que vous n'aurez pas servi chaque client (pour le fonds de cautionnement, cela impliquerait d'avoir la capacité de payer la caution de chaque personne qui en a besoin), certains, "aléatoirement" choisis, ne recevront pas ce service. Une RCT ne fait que rendre ce processus aléatoire plus explicite. Par exemple, supposons qu'on ne dispose de ressources que pour doter les tribunaux en personnel cinq jours par semaine. Randomisez les jours où il y a du personnel. Ou bien dites que vous n'avez pas le temps de rencontrer tous les accusés quand le personnel est présent. Commencez par rencontrer les accusés ayant des numéros de dossier impairs puis, si vous avez le temps, passez aux

5. Par exemple, le Conseil indien pour la recherche médicale manifesta son inquiétude relative au respect des règles éthiques par les études expérimentales (CHATTERJEE, 2008). MUDUR (2005 : 1044) cite le Pr. Falguni Sen, de la Fordham University à New York, disant qu'« Étant donné la vulnérabilité des patients pauvres et peu éduqués, l'Inde a encore un long chemin à parcourir pour assurer une protection adéquate des sujets humains ».

6. Nous remercions Tim Ogden pour nous avoir signalé cette conversation sur Twitter.

> numéros pairs. Ces deux méthodes sont des RCT qui peuvent produire des données très précieuses ! Car vous avez beau penser que vous savez que votre programme est extrêmement efficace, vous n'en avez en fait aucune idée. Accédez-vous aux clients qui ont le plus besoin de vous ? Ceux qui seraient en souffrance si vous ne leur apportiez pas votre aide ? » (Pr. Stevenson, 14 mars 2019).

Reformulé en termes d'équipoise, l'argument de MBF stipule que l'attribution aléatoire d'une caution au groupe traité et sa non-attribution au groupe contrôle violent le principe d'équipoise parce que le groupe contrôle est désavantagé. La Pr. Stevenson réfute cet argument au motif que les connaissances de terrain ne suffisent pas à évaluer avec certitude la supériorité du traitement (« vous avez beau penser que vous savez […] vous n'en avez en fait aucune idée »). Elle accorde à la recherche d'une validation scientifique rigoureuse une supériorité par rapport à la crainte que l'équipoise ne soit pas respectée[7]. Ce faisant, elle utilise le raisonnement typique des économistes du développement, à savoir que la vie réelle est déjà injuste/incertaine pour les sujets des RCT (« Tant que vous n'aurez pas servi chaque client »). L'idée ainsi mise en exergue est que la RCT est importante pour la science et pour les générations futures.

Il est aussi possible que les économistes considèrent leurs propres études expérimentales comme bénignes, en ce sens que les organisateurs n'ont pas à se soucier des conséquences potentiellement négatives pour les membres du groupe contrôle qui ne reçoivent pas les bénéfices du traitement testé. Par exemple, on jugera que le fait de ne pas octroyer à un agriculteur le prêt dont il pourrait avoir besoin ne constitue pas un préjudice important. En termes économiques, une situation où 50 % des agriculteurs reçoivent le prêt et 50 % ne le reçoivent pas, améliore la situation au sens de Pareto[8] par rapport au *statu quo*, où aucun prêt n'est octroyé.

Cependant, les gens vivent dans des communautés où s'appliquent des normes sociales et les changements apportés à la vie de certains individus peuvent avoir des conséquences notables pour l'ensemble de la société[9]. Une illustration de ce phénomène disruptif est donnée par l'intervention de prévention du syndrome d'immunodéficience acquise (SIDA), qui implique de tester tous les participants au virus de l'immunodéficience humaine (VIH). Dans ce cas, et dans bien d'autres, l'étude expérimentale peut avoir des effets majeurs, particulièrement observables lorsque les normes sociales sont perturbées (Morvant-Roux *et al.*,

7. Dans leur chapitre, Garchitorena *et al.* (2018) mentionnent que « de nombreuses RCT sont effectuées dans le but de confirmer les résultats d'études observationnelles ».

8. C'est-à-dire dans le sens où il n'est pas possible d'améliorer la situation d'une personne sans dégrader celle d'un autre au moins.

9. Afin de limiter la portée de ce problème, la randomisation est parfois appliquée de façon groupée (par *cluster*), au niveau d'une communauté. Les RCT groupées peuvent néanmoins aussi être sujettes à la préoccupation de l'équipoise. Par exemple, dans certaines RCT qui explorent les effets de la distribution de filets de protection de lits imprégnés d'insecticide, un seul groupe reçoit ces filets gratuitement (Müller *et al.*, 2008 ; Tarozzi *et al.*, 2014).

2014). Des effets de contrecoup sont typiquement observés dans le contexte d'interventions visant à favoriser l'émancipation des femmes dans les sociétés patriarcales (Schuler *et al.*, 2018). Comme le théorise Rabin (1993), l'équité économique peut avoir des conséquences sur le bien-être social. Les gens peuvent trouver acceptable de ne rien recevoir dès lors que le même sort est réservé à tous, mais s'y opposer lorsque certains, désignés au hasard, sont récompensés d'une façon perçue comme injuste. De plus, si le traitement testé n'est censé introduire qu'une petite différence par rapport à ceux qui ne reçoivent rien, alors peut-être que ce traitement ne mérite pas d'être étudié avec un plan expérimental aussi lourd (et cher) qu'une RCT. Le recours aux RCT suppose en effet que les enjeux soient suffisamment importants pour justifier du déploiement d'un plan expérimental sophistiqué et coûteux. En clair, ignorer, au moins partiellement, l'équipoise implique que l'étude concernée s'apparente soit à un gaspillage d'argent, soit à une source de « malaise moral » (Baele, 1913 : 4).

Contrairement aux médecins, les économistes du développement n'ont pas l'habitude de voir l'éthique interférer avec leurs méthodes de recherche. En général, l'éthique ne figure pas au rang des préoccupations des *randomistas*, et elle constitue encore moins un but dans les plans expérimentaux (Barrett et Carter, 2010). Leurs objectifs résident ailleurs, souvent dans d'ambitieuses politiques publiques pour « résoudre la pauvreté » (Karlan et Appel, 2011) et dans la mise à l'épreuve de théories économiques. Banerjee et Duflo (2009 : 156) considèrent les expériences comme « un puissant outil pour tester les théories ». Ces deux types de préoccupations divergent de celles des chercheurs en médecine.

Les recommandations de politique économique sont sans doute plus proches de l'intérêt du médecin pour la situation de son patient. Il y a toutefois une nuance notable, puisque la politique économique est pensée comme un traitement général pour un grand ensemble d'individus, dont certaines ne cherchent même pas à se faire traiter. Le problème de la participation au traitement fréquemment rencontré par les *randomistas* se pose lorsque les participants visés ne sont pas intéressés par le traitement. La réaction type de l'investigateur consiste alors à susciter la participation à l'étude par quelque sorte d'encouragement (White, 2013), qui peut d'ailleurs avoir des conséquences délétères sur les résultats de l'étude. En revanche, les RCT médicales font appel à des investigateurs locaux, de terrain, pour évaluer l'état de santé des patients qui seront recrutés dans l'étude. Cette procédure empêche les chercheurs d'inclure des participants non malades, tant dans le groupe traité que dans le groupe contrôle. Pour les économistes, procéder de la même manière, sans causer de préjudice, reviendrait à identifier *ex ante* les besoins de tous les participants et donc à construire deux options d'attrait égal (incluant éventuellement une certaine forme de rémunération) et à les proposer aléatoirement au groupe traité et au groupe contrôle. D'où l'équipoise.

Le contraste entre les objectifs de la politique économique et ceux de la théorie rappelle ce que Weijer *et al.* (2000) présentent comme une controverse entre le Royaume-Uni et les États-Unis ou, plus largement, comme l'arbitrage entre la

science et le patient. Or, en économie expérimentale, il n'y a pas de consensus sur une option intermédiaire qui offrirait au groupe contrôle l'équivalent du « traitement standard plus placebo », afin de ne pas nuire aux personnes qui ne reçoivent pas le traitement expérimental. En fait, la plupart des RCT économiques vont dans la direction opposée en testant des traitements dont on attend des résultats généralement positifs. Ceci explique que les chercheurs et praticiens de la microfinance ont été fort déçus par les résultats des RCT montrant des impacts fort modestes (Duvendack *et al.*, 2011). Si ces expériences avaient été soumises au principe de l'équipoise, la déception aurait été moindre au vu des résultats mitigés des RCT. Mais, paradoxalement, la modestie des impacts obtenus par ces RCT pourrait être présentée comme la confirmation *a posteriori* d'une certaine forme d'équipoise. Cependant, ce domaine de recherche pose le problème récurrent suivant : les impacts économiques examinés varient d'une publication à l'autre. Il en résulte qu'on teste des éléments disparates relatifs au bien-être des sujets. Par exemple, l'étude d'Attanasio *et al.* (2015) repose sur un large éventail de mesures d'impact comme les accroissements de l'entrepreneuriat, de la scolarisation, de la consommation et des taux de remboursement des prêts. Cette hétérogénéité des mesures d'impact rend la mise en œuvre de l'équipoise particulièrement complexe dans les sciences sociales[10].

Le peu de place accordée à l'éthique dans les RCT économiques rappelle l'absence de réactions observée autrefois à la suite d'expérimentations médicales non éthiques pratiquées sur des populations défavorisées. Dans l'étude Tuskegee (1932-1972), organisée par le service public américain de la santé, le groupe étudié était composé d'hommes afro-américains pauvres atteints d'une syphilis non soignée (Caplan, 2001). La justification avancée par les expérimentateurs pour ne donner aucun traitement à ces patients était que ces hommes pauvres n'auraient de toute façon pas eu les moyens de se payer le traitement (Angell, 1997). Cet argument peut sembler familier à ceux qui ont interrogé des *randomistas* sur l'injustice envers les membres de leurs groupes contrôle. Il faut noter que l'étude Tuskegee a été interrompue à la suite d'une intervention des médias – le *Washington Star* et le *New York Times* – qui a embarrassé l'administration Nixon (Angell, 1997).

Lurie et Wolfe (1997) soulignent les différences entre les conceptions des RCT médicales mises en place aux États-Unis et dans les pays en développement[11] pour tester de nouveaux médicaments destinés à sauver la vie des nouveau-nés de femmes infectées par le sida. Deux expériences organisées aux États-Unis, où les groupes étudiés avaient accès à des médicaments antirétroviraux, sont ainsi opposées aux études réalisées dans des pays en développement, où la plupart

10. Le problème des mesures d'impact multiples dépasse le cadre de la microfinance. Ainsi, une étude récente de Schilbach (2019) sur l'impact des dispositifs d'engagement sur la consommation d'alcool des conducteurs de pousse-pousse en Inde utilise des mesures d'impact telles que la consommation d'alcool, la sobriété en journée, la productivité, les gains et l'épargne. Quoi qu'il en soit, l'attribution aléatoire d'incitations à la sobriété pose question sur le plan éthique.

11. Côte d'Ivoire, Ouganda, Tanzanie, Afrique du Sud, Malawi, Éthiopie, Burkina Faso, Zimbabwe, Kenya et République dominicaine.

des patientes n'avaient pas accès à ces médicaments, probablement pour des raisons financières. LURIE et WOLFE (1997) rapportent également l'anecdote d'un chercheur de Harvard qui a demandé aux National Institutes of Health (NIH) américains de financer une étude éthiquement bien conçue – avec un groupe contrôle traité activement – en Thaïlande et dont la réponse des NIH invitait ses auteurs à réduire les coûts en réalisant plutôt une étude contre placebo. Ce n'est qu'après que le directeur du comité des sujets humains de Harvard a signalé qu'en l'occurrence l'usage du seul placebo était contraire à l'éthique, que les NIH ont finalement admis l'argument. Au-delà de l'anecdote, PETRYNA (2009) observe que la part des RCT médicales organisées dans les pays émergents[12] a augmenté, passant de 10 % en 1991 à 40 % en 2005. L'auteure s'interroge sur le caractère abusif de la délocalisation des essais cliniques, dont les sociétés pharmaceutiques peuvent se servir pour inciter les médecins des pays émergents à prescrire des médicaments coûteux, et ainsi nuire à la fourniture de traitements abordables. De toute évidence, la toute grande majorité des économistes expérimentaux n'ont pas de telles motivations lucratives. Pourtant, les RCT sont coûteuses à mettre en œuvre et privent ainsi de fonds d'autres études, souvent moins onéreuses. De plus, des expériences médicales passées dans les pays en développement ont mis en évidence le risque réputationnel associé aux études réalisées sur les populations pauvres. Ces populations sont faciles à exploiter parce qu'elles ignorent généralement leurs droits à une information complète sur le plan expérimental et, ensuite, à formuler soit un consentement éclairé, soit un refus.

Une autre différence notable entre les études médicales et les études économiques réside dans l'intensité – par opposition à la nature – de l'impact recherché, puisque les perturbations économiques qui viennent altérer les structures et traditions existantes sont jugées mineures par rapport aux essais de médicaments. En outre, la plupart des RCT réalisées par les économistes dans les pays en développement visent, d'une manière ou d'une autre, à aider ces pays. On ne peut donc pas soupçonner ces études de poursuivre un but commercial ou de se servir des pays en développement comme laboratoire pour tester des traitements destinés aux économies développées. Mais comme les RCT sans équipoise attribuent des situations supposées favorables à des personnes choisies au hasard, on peut leur opposer un argument souvent avancé en médecine lorsque les traitements sont rationnés : ne conviendrait-il pas que le traitement soit offert à ceux qui en ont le plus besoin ? Dans cette optique, BAELE (2013 : 19) affirme que « la randomisation viole également le principe moral prioritaire consistant à assurer un certain niveau de bien-être à la sous-population la plus défavorisée avant de penser soit à maximiser en termes absolus la richesse de la population (version conséquentialiste), soit à garantir la liberté individuelle (version libérale). » Par conséquent, même bénins, les traitements économiques ne justifient pas que l'on néglige l'exigence éthique d'équipoise envers le groupe contrôle.

12. L'auteur mentionne des exemples relatifs à l'Europe de l'Est et au Brésil.

Si la profession économique adopte l'exigence d'équipoise, sa mise en œuvre imposera de fortes contraintes qui devraient conduire à ce que le groupe traité et celui de contrôle reçoivent des traitements *a priori* qualitativement similaires, avec validation par avis d'experts. Cependant, la préférence des bailleurs de fonds pour un impact prometteur peut exercer une pression opposée à celle de l'équipoise (Ravallion, chap 1, ce volume). Dans l'ensemble, même si la mise en œuvre pratique de cette procédure peut se révéler fastidieuse, elle permettra au moins d'exclure les comportements ouvertement contraires à l'éthique.

En outre, les exemples médicaux décrits dans cette section démontrent que, même dans les sciences médicales, où la déclaration d'Helsinki et l'exigence d'équipoise sont supposées faire fonction de boussoles morales, le relativisme éthique a la vie dure. Les chercheurs qui effectuent des RCT dans des pays en développement ont tendance à avancer un argument financier pour abaisser les normes de soins, au motif que la plupart des gens et/ou leurs gouvernements ne peuvent pas se payer les meilleurs traitements. Rejetant cet argument, l'article éditorial d'Angell (1997 : 848) a annoncé que le prestigieux *New England Journal of Medicine* avait décidé de ne plus publier d'articles basés sur « recherches non éthiques, et ce quelle que soit leur valeur scientifique ». Les revues économiques suivront-elles la même voie ? Pour le moment, rien ne le laisse présager.

Conclusion

Le temps est venu de demander aux économistes du développement adeptes des RCT et autres *randomistas* de peser les conséquences éthiques de leurs actes à l'aune des objectifs ultimes des études menées. Comme ce fut le cas au sein de la communauté médicale des années 1980, de plus en plus de chercheurs remettent en question le paradigme des RCT vues comme l'« étalon-or » de la recherche (Cartwright, 2007 ; Deaton, 2010a ; Bédécarrats *et al.*, 2019b). Mais, alors même que les RCT peuvent avoir une incidence significative sur des vies humaines, les critiques fondées sur l'éthique figurent, jusqu'à présent, parmi les moins virulentes. Cette situation est probablement liée aux bonnes intentions sous-jacentes, comme celle visant à aider les pauvres dans la prévention du paludisme (Cohen et Dupas, 2010) et dans l'installation sanitaire (Duflo *et al.*, 2015b). En même temps, comme l'ont dit Baldassarri et Abascal (2017 : 62), les « expérimentateurs de terrain "jouent à Dieu", en intervenant dans la vie des gens de manière conséquente. » Souvent, une connaissance préalable du terrain permet de prévoir les résultats négatifs avec un certain degré de confiance. Dans de tels cas, l'absence d'équipoise combinée aux risques spécifiques auxquels sont exposées les personnes défavorisées résulte soit de l'indifférence, soit d'une expérience de terrain insuffisante chez les expérimentateurs. D'un point de vue éthique, cette situation est même pire que celle des études attribuant au

hasard des traitements présumés favorables. Mais toutes les deux peuvent laisser des traces indélébiles, non seulement au niveau individuel, mais aussi sur les relations interpersonnelles au sein des communautés.

À l'inverse, les arguments dénonçant l'iniquité de plans expérimentaux négligent les bénéfices de long terme apportés par les connaissances nouvelles que les RCT peuvent générer. Même dans la sphère médicale, certains auteurs critiquent la notion d'équipoise ou proposent de la modifier (Fries et Krishnan, 2004 ; Ubel et Silbergleit, 2011). Ils déplorent que l'on ne tienne pas suffisamment compte des avantages futurs, qui sont, de fait, ignorés en regard de l'obligation thérapeutique des médecins envers leurs patients. Veatch (2007 : 182) affirme que « ce n'est pas quelque équipoise qui est moralement fondamentale ; c'est le fait que les sujets potentiels consentent à être randomisés sans être indûment contraints, manipulés ou exploités ». Miller et Brody (2007 : 153) estiment pour leur part « que les principes éthiques régissant la thérapie médicale sont différents de ceux régissant la recherche clinique ». Les auteurs soulignent l'importance des rendements attendus de l'étude, y voyant là une raison légitime de s'écarter de l'exigence d'équipoise. Cette logique revient à accepter de sacrifier le bien-être de certains individus (généralement ceux du groupe contrôle) pour le bien supérieur de la société et des générations futures. Elle est toutefois atténuée par le double aveugle, qui rend le sacrifice aléatoire, plutôt que certain, et le disperse ainsi sur tous les participants à l'étude. Les RCT économiques n'ont généralement pas recours au double aveugle, de sorte qu'elles peuvent difficilement invoquer cette excuse commode. En tout état de cause, les études scientifiques traitant injustement ou infligeant des sacrifices à des individus, tout spécialement s'ils sont déjà défavorisés, seront toujours éthiquement contestables. Et, comme les études moralement discutables sont préférentiellement menées dans des pays à faible protection juridique, l'imposition de l'équipoise, sous une forme ou une autre, à l'expérimentation économique devrait dépasser la rhétorique.

Remerciements

Les auteurs remercient Cécile Abramowicz, Britta Augsburg, Marie Brière, Andres Garchitorena, Marek Hudon, Marc Labie, Jonathan Morduch, Tim Ogden, Martin Ravallion, les participants à l'atelier de l'AFD *Randomized Control Trials in the Field of Development: The Gold Standard Revisited* (Paris, mars 2019) et les trois éditeurs du livre, Florent Bédécarrats, Isabelle Guérin et François Roubaud, pour leurs précieux commentaires et les discussions passionnantes.

Chapitre 11

Utilisation des *a priori* dans les protocoles expérimentaux

Que laisse-t-on de côté en les ignorant ?

Eva Vivalt

Introduction

Les mérites relatifs des évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT) ont fait l'objet de nombreux débats. Dans ce chapitre, je laisse de côté cette vaste discussion pour me concentrer sur une question pertinente, étroite et peu étudiée : dans quelle mesure l'exploitation des *a priori* peut améliorer la conception des études ? Par *a priori*, j'entends les croyances préalables qu'ont les décideurs sur les effets d'un programme particulier. Par exemple, certains décideurs politiques peuvent penser qu'un programme de transferts monétaires sans conditionnalité a de fortes chances d'avoir un impact important sur les résultats éducatifs. Si le processus décisionnel et les *a priori* étaient connus, les chercheurs seraient en mesure de mieux cibler leurs évaluations d'impact. Dans certains cas, il peut y avoir de nombreux bras (*arms*) de traitement à tester, et les *a priori* peuvent indiquer les bras qui doivent être mis en œuvre et évalués. De même, les bras de traitement peuvent être déterminés, mais sans que l'on sache quels indicateurs de résultats privilégier. Plus particulièrement, les chercheurs sont souvent amenés à décider des résultats à inclure dans les enquêtes intermédiaires ou finales en tenant compte des contraintes de temps. Les *a priori* des décideurs pourraient renseigner sur les indicateurs de résultats à collecter le plus fréquemment afin d'obtenir une meilleure puissance statistique, toutes choses égales par ailleurs. On peut également imaginer que, selon les *a priori* des décideurs, la taille de l'échantillon de l'étude doive être plus ou moins grande pour que l'étude produise des preuves convaincantes, de sorte que l'utilisation d'*a priori* pourrait renforcer l'efficacité d'une évaluation d'impact pour informer la politique publique.

Tant les RCT que les études appliquant des méthodes quasi expérimentales (désignées ci-après comme « non-RCT ») pourraient en principe tirer profit des *a priori*. Cependant, ces méthodes interagissent également avec celle de l'*a priori*. Dans la mesure où l'on peut considérer que les RCT sont motivées par le désir de convaincre un décideur parmi les plus adverses qui soient (Banerjee *et al.*, 2017a), l'exploitation d'un ensemble *spécifique* d'*a priori* semble philosophiquement plus en accord avec les non-RCT, c'est-à-dire que nous nous inquiéterions moins d'avoir un public parmi les plus défavorables si l'étude était conçue pour convaincre un public spécifique, dont les *a priori* sont connus. En outre, seules des non-RCT peuvent assigner de manière déterministe des participants à des groupes de traitement, ce qui est nécessaire pour produire certains des bénéfices attendus de l'exploitation des *a priori*. D'autre part, les résultats des non-RCT peuvent susciter un certain scepticisme justifiable par rapport à ceux des RCT, au motif que les chercheurs peuvent s'engager consciemment ou inconsciemment dans la recherche de spécifications avec des non-RCT. Il y a recherche de spécifications, ou « *p-hacking* », lorsque les chercheurs pratiquent de nombreux tests statistiques (par exemple en effectuant de multiples régressions avec différentes variables de contrôle) et font partialement état des résultats significatifs. Cela conduit à des inférences erronées. Une recherche de spécifications plus poussée a été observée à plusieurs reprises dans des non-RCT utilisant des tests de significativité classiques (Brodeur *et al.*, 2016 ; 2018 ; Vivalt, 2019)[1]. La collecte et l'exploitation des *a priori* étant encore une nouvelle approche en économie, il pourrait être plus intéressant de s'en servir pour les RCT, car les choix pratiqués au cours du processus gagneraient en transparence. Le reste de ce chapitre développera ces divers points, après avoir décrit plus en détail ce que l'on entend par « croyances *a priori* » et comment celles-ci peuvent être explicitées.

Explicitation et utilisation des *a priori*

De plus en plus de chercheurs procèdent *ex ante* à une collecte des *a priori* sur les effets que leurs études vont mettre en évidence. Par exemple, une équipe chargée d'évaluer l'impact d'un programme de transferts monétaires conditionnels – dans le cadre duquel des ménages reçoivent de l'argent en échange de la scolarisation de leurs enfants en âge d'aller à l'école – pourrait demander

1. Nous considérons normalement la recherche de spécifications comme un phénomène qui se produit lors de l'utilisation de tests classiques (« fréquentistes »), par exemple pour vérifier si une relation semble significative au niveau de 5 %. Cependant, il est possible que, même dans un cadre bayésien, si les chercheurs connaissaient les *a priori* des décideurs, ils pourraient mener une recherche de spécifications soit pour appuyer, soit pour fragiliser les *a priori* des décideurs et influencer les décisions politiques. Par conséquent, même aux yeux d'un bayésien, les RCT peuvent sembler plus crédibles.

à d'autres d'essayer de deviner l'effet qu'aurait le programme sur les taux de scolarisation. Dans le cadre de ce processus, l'équipe de chercheurs commencerait par décrire le programme et son contexte avec force détails afin de susciter des suppositions plus précises.

Il y a plusieurs raisons à vouloir recueillir ces croyances préalables. Tout d'abord, ces prévisions peuvent être intéressantes en soi, car on peut en déduire si certains sous-groupes de répondants formulent des prévisions plus précises. Par exemple, certains travaux suggèrent que les décideurs politiques ont tendance à avoir des croyances plus optimistes que les chercheurs (VIVALT et COVILLE, 2016 ; CASEY *et al.*, 2018). Au fil du temps, nous pouvons apprendre à mieux cerner les conditions dans lesquelles les individus établissent de meilleures prévisions ou à identifier les individus qui sont plus performants dans cet exercice. Nous pourrions également apprendre à « dé-biaiser » les prévisions autant que faire se peut par la modélisation ou une approche d'apprentissage automatique. À l'issue du processus, il nous reste un résultat potentiellement intéressant : des prévisions avec un certain contenu informationnel.

Ces prévisions peuvent être importantes pour l'élaboration des politiques, car nous ne pourrons jamais mener autant d'études que nous le souhaiterions. En l'absence de preuves empiriques issues d'études académiques, des prévisions débiaisées peuvent aider les décideurs à déterminer les interventions à conduire. Notez que nous n'avons pas besoin que les prévisions soient toujours précises pour être exploitables à cet effet. Il suffit de corréler les prévisions avec les effets des interventions pour qu'elles soient utiles, même si des erreurs sont encore commises.

De manière plus triviale, chaque équipe de recherche est incitée à titre privé à collecter les croyances prévalantes *ex ante :* cela peut être rentable en termes de publication. En particulier, les revues académiques privilégient souvent les résultats qui sont statistiquement significatifs. En recueillant des *a priori ex ante*, les chercheurs bénéficient d'une certaine sécurité par rapport à l'état des choses dans lequel ils obtiennent des résultats « nuls » (c'est-à-dire dès lors qu'ils constatent que le programme a eu un effet nul). Dans de tels cas, les *a priori* leur permettront parfois d'arguer de manière crédible que ces résultats nuls étaient inattendus et qu'ils gardent donc un intérêt académique[2].

Le reste de ce chapitre se concentrera sur un autre avantage de la collecte de croyances préalables *ex ante* sur les effets des interventions : leurs bénéfices potentiels pour concevoir les protocoles d'expérimentation.

2. À long terme, si la collecte d'*a priori* prenait son essor, on pourrait imaginer que le fait de juger de l'originalité et de l'importance de résultats de recherche recourant aux croyances préalables pourrait supplanter les comparaisons par rapport à l'absence d'effets et que le biais de publication pourrait se déplacer vers les études dont les résultats seraient « originaux ».

Utilisation des *a priori* dans les protocoles d'expérimentation

Les *a priori* peuvent aider à concevoir un protocole expérimental de plusieurs manières. Premièrement, on peut modifier l'assignation d'un échantillon potentiel à différents groupes de traitement ou collecter différents indicateurs de résultats en fonction de la valeur de l'information propre à chaque assignation. Or, les *a priori* pourraient également améliorer un protocole expérimental par une assignation déterministe, un fait connu depuis très longtemps (voir, par exemple, les débats entre FISHER [1960], et GOSSET [1937]). Pour le contexte, il peut être utile d'examiner un cas tiré de BANERJEE *et al.* (2017a), dans lequel une expérimentation de chèques scolarité est prévue pour aider un haut fonctionnaire en charge de l'éducation à décider s'il doit ou non utiliser ce type de mode de financement. Le haut fonctionnaire en charge de l'éducation estime que le fait qu'un élève soit issu d'une famille riche ou pauvre est un facteur déterminant de la réussite scolaire, mais il est également ouvert à l'idée que la qualité de l'école puisse être très importante, de sorte que même un élève pauvre pourrait apprendre davantage dans une école privée. Le haut fonctionnaire est autorisé à affecter un élève à une école privée. D'un point de vue classique, il est impossible d'apprendre quelque chose de significatif d'une expérience à partir d'une seule observation. Cependant, pour un bayésien, il y a toujours quelque chose à apprendre. En effet, si un enfant pauvre était assigné à l'école privée et obtenait aux examens standardisés des résultats meilleurs que ce que le haut fonctionnaire en charge de l'éducation pouvait imaginer d'un enfant pauvre, cela serait informatif et le haut fonctionnaire serait amené à revoir ses croyances.

Les *a priori* du décideur jouent un rôle important ici. En raison de ces *a priori*, le fait d'assigner un enfant riche à cette école privée ne serait pas aussi informatif que si c'était un enfant pauvre. De même, si une place dans une école publique se libérait, il ne serait pas logique d'y envoyer un enfant pauvre du point de vue de la valeur de l'information qu'on pourrait en retirer. Ces exemples montrent qu'un bayésien devrait assigner les sujets aux groupes de traitement et de contrôle de manière déterministe, plutôt qu'aléatoire. Des arguments similaires ont été avancés ailleurs (KASY, 2016).

BANERJEE *et al.* (2017a) montrent ensuite que la randomisation peut aider à convaincre un public adverse. Par exemple, supposons qu'au lieu d'avoir un seul haut fonctionnaire en charge de l'éducation avec un seul ensemble de croyances, il y a un groupe de hauts fonctionnaires pressentis pour ce poste, avec tout l'éventail de croyances possibles sur la relation entre la réussite scolaire, le fait qu'un élève soit riche ou pauvre et le fait qu'il fréquente une école privée ou publique[3]. BANERJEE *et al.* (2017a) nous demandent d'examiner le cas du

3. Par exemple, il pourrait y avoir un renouvellement des hauts fonctionnaires en charge de l'éducation, ce qui induirait une certaine variation des croyances.

choix, parmi les candidats potentiels, du haut fonctionnaire ayant l'*a priori* qui maximise les chances de mise en œuvre de la mauvaise politique. Dans ce monde « adverse », il n'y a plus aucun intérêt à envoyer un enfant pauvre dans une école privée et un enfant riche dans une école publique. De même, si un décideur n'avait pas confiance en son propre *a priori* et se montrait adverse à l'ambiguïté, la randomisation pourrait se révéler utile en tant que stratégie mixte.

Les enseignements de Banerjee *et al.* (2017a) sont excellents et expliquent pourquoi les chercheurs qui essayent de convaincre les évaluateurs d'articles des revues scientifiques préfèrent les RCT – les référés constituent certainement le public le plus adverse qui soit ! On comprend également pourquoi les entreprises qui mènent des expériences à usage interne, souvent avec des échantillons de petite taille, sont moins susceptibles de procéder à une randomisation : avec des échantillons plus petits, les gains en puissance statistique de l'assignation déterministe sont plus importants, et ces entreprises peuvent concevoir leur évaluation de manière à prendre une décision spécifique à l'appui de certains *a priori* particuliers. Or, les décideurs politiques semblent se situer quelque part entre ces deux extrêmes. Il est peu probable qu'ils soient les plus adverses qui soient, mais ils peuvent avoir des *a priori* qui se situent dans une fourchette restreinte. Cependant, les décideurs politiques ne ressemblent pas tout à fait à des entreprises, car ils peuvent être en mesure d'exploiter de plus gros échantillons pour leurs essais et être relativement incertains quant à l'ensemble exact des *a priori* qui seront pertinents à l'avenir, par exemple s'il y a une élection ou une autre raison de renouvellement des élus avant que les résultats de l'évaluation ne soient connus. De même, les décideurs politiques peuvent avoir plus souvent tendance à voir dans l'évaluation un bien public pour d'autres décideurs politiques. Ils peuvent en outre avoir à cœur de convaincre leurs électeurs, lesquels sont susceptibles d'avoir des croyances préalables très diverses, sur les bienfaits du programme social en question. Dans la mesure où les *a priori* des autres sont inconnus ou adverses, les RCT peuvent mieux convenir à un décideur politique défavorable à l'ambiguïté.

On pourrait également faire valoir que les chercheurs devraient opter pour les RCT dès lors que des biais sont susceptibles de se glisser dans des études quasi expérimentales. Brodeur *et al.* (2016 ; 2018) ont constaté que les résultats des RCT et des expériences de laboratoire sont moins souvent dans la catégorie « juste au-dessus » du seuil classique de 5 % de significativité que d'autres types d'études dont les résultats se situent « juste en dessous »[4] ; avec un autre ensemble de données, Vivalt (2019) a constaté qu'il y avait 17 % de moins de RCT juste au-dessus du seuil par comparaison aux études non-RCT. C'est là une preuve convaincante que les RCT présentent moins de signes de recherche de spécifications que les non-RCT lorsqu'il est question de tests de significativité classiques. Un bayésien ne se soucierait pas des tests de significativité classiques ; cependant,

4. Brodeur *et al.* (2018) ont toutefois constaté que l'un des types de protocoles quasi expérimentaux qu'ils ont examinés, le protocole de régression sur discontinuités, obtenait de meilleurs résultats que les RCT.

il existe d'autres façons de fausser des résultats. En particulier, la recherche de spécifications qui amplifie la significativité des tests classiques tendra également à exagérer l'ampleur de l'effet estimé. Ainsi, si la recherche est pratiquée à la fois à des fins de publication, pour laquelle les tests classiques importent, et pour informer un choix politique, dont les décideurs pourraient être bayésiens, on peut toujours s'interroger sur l'ampleur des effets rapportés.

Les RCT peuvent présenter moins de recherches de spécifications classiques pour deux types de raisons, chacune ayant des implications différentes pour les bayésiens. Premièrement, les chercheurs peuvent être moins incités à rechercher des spécifications si leurs études sont plus faciles à publier. Deuxièmement, il peut y avoir quelque chose dans les RCT qui, intrinsèquement, rend la recherche de spécifications moins probable, par exemple le fait que les chercheurs utilisent plus souvent des plans d'analyse préalables qui rendent plus difficiles la recherche de spécifications, le fait qu'il soit potentiellement plus difficile de justifier d'actions comme l'inclusion de diverses variables de contrôle puisque, en principe, la randomisation devrait conduire à un équilibre des covariables dans les grands échantillons, ou le fait que les RCT induisent des connotations de rigueur qui incitent les auteurs à ne pas se livrer à de mauvaises pratiques. Le premier type de raisons peut ne pas se poser dans le contexte du processus décisionnel bayésien, alors même que, si les chercheurs voulaient influencer une décision politique, ils pourraient être pareillement incités à fausser des résultats. Dans la mesure où le deuxième type de raisons favorise la recherche de spécifications dans une perspective classique, il demeurerait pertinent dans une conception bayésienne.

On peut supposer que toute recherche de spécifications ne ferait qu'exagérer les effets sans rien changer aux signes de ces effets. Ainsi, si l'on était bayésien, la bonne conclusion ne serait pas de faire preuve d'une défiance totale vis-à-vis des résultats non-RCT, mais peut-être de réduire les estimations par un « facteur d'exagération » type[5]. Si l'on croit que les non-RCT sont biaisées, on peut ironiquement être plus défavorable à leurs résultats. Mais ces biais ne doivent pas non plus signifier l'abandon total des non-RCT. De meilleures normes et dispositifs de contrôle, tels que des plans d'analyse préalables, pourraient contribuer à résoudre ce problème.

La problématique ci-dessus s'est concentrée sur un type particulier de biais qui peut altérer les résultats d'une étude. La question de savoir si les non-RCT sont plus biaisées que les RCT s'avère plus compliquée que ne le suggère cette explication. En particulier, dans d'autres chapitres, il est argué que des RCT peuvent être réalisées dans des lieux déterminés, ce qui entraîne un biais de sélection des sites, ou qu'elles peuvent avoir des échantillons plus petits que des non-RCT, ce qui entraîne des intervalles de confiance plus importants pour leurs résultats, ainsi que des prédictions basées sur les résultats affichant des erreurs quadratiques

5. Par exemple, Gelman et Tuerlinckx (2000) et Gelman et Carlin (2014) introduisent des erreurs de « type M » (pour « magnitude »), qui pourraient indiquer à quel point un résultat est susceptible d'être exagéré.

plus importantes (en raison du compromis biais-variance). Je fais abstraction de ces problèmes potentiels pour deux raisons. Premièrement, dans d'autres travaux portant sur un échantillon de 635 études de 20 interventions dans le domaine du développement international, je ne parviens pas à rejeter l'hypothèse nulle selon laquelle les tailles des effets estimées par les RCT et les non-RCT sont les mêmes (Vivalt, 2020b). Deuxièmement, il existe de nombreuses situations dans lesquelles les coûts des RCT et des non-RCT sont identiques, de sorte que les RCT n'auront pas nécessairement des échantillons plus petits. Quoi qu'il en soit, le compromis biais-variance est une question extrêmement sous-estimée en économie. Dans un monde idéal, un décideur politique ne devrait pas seulement se préoccuper de l'éventuel biais de l'estimation ponctuelle, mais aussi de la précision des estimations et de l'erreur totale de prévision. Certains éléments indiquent que les décideurs politiques souffrent plutôt d'une « incurie de la variance », qui comprend mal ou ignore les intervalles de confiance (Vivalt et Coville, 2016).

Estimations approximatives des avantages de l'utilisation des *a priori*

La question demeure : quels sont les avantages de l'utilisation des *a priori* dans les protocoles d'étude ? Je conduis quelques simulations simples qui produisent des estimations approximatives. Dans cette section, j'étudie les avantages potentiels qui sont jugés communs aux RCT et aux non-RCT pour souligner le fait que l'exploitation des croyances préalables, comme de nombreuses autres questions méthodologiques majeures, peut être importante aussi bien pour les RCT que pour les non-RCT.

Supposons qu'il y ait deux interventions – par exemple un programme de transferts monétaires conditionnels et un programme de restauration scolaire – et qu'un décideur politique doive décider laquelle mettre en œuvre pour améliorer les taux de scolarisation. Afin de prendre cette décision, le décideur politique demande à un chercheur de planifier une évaluation d'impact d'un pilote de l'une des deux interventions. Si le chercheur ne connaît pas les *a priori* du décideur politique, il ne saura pas sur quelle intervention pratiquer l'évaluation d'impact et pourrait choisir la mauvaise. Par exemple, supposons que le décideur politique soit très incertain des effets du programme de restauration scolaire, mais assez certain des effets du programme de transferts monétaires conditionnels. Il serait alors logique de procéder à l'évaluation d'impact sur le programme de restauration scolaire, car il est peu probable que le responsable politique change d'avis sur les effets du programme de transferts monétaires conditionnels.

Je produis des estimations approximatives des avantages de la prise en compte des *a priori* à l'aide de l'algorithme suivant : en premier lieu, je spécifie un *a priori* pour chaque intervention. Par souci de simplicité, je suppose que les *a priori* sont normalement distribués, ce qui nécessite de spécifier simplement

une moyenne et un écart-type. Je suppose que l'écart-type de la variable de résultat est le même pour chaque intervention et que, en l'absence d'*a priori*, chaque intervention aurait les mêmes chances d'être sélectionnée pour l'évaluation d'impact. Je détermine ensuite la valeur de l'information, du point de vue du décideur politique, pour faire une évaluation de l'impact de chaque intervention compte tenu des *a priori* qui ont été établis. La valeur de l'information est une construction désignant la probabilité qu'une évaluation d'impact vienne changer la décision qui a été prise, le décideur préférant toujours l'intervention présentant *a posteriori* l'effet moyen le plus élevé, multiplié par le bénéfice attendu de cette décision, c'est-à-dire la différence entre les moyennes réelles des effets des deux interventions. Il convient de souligner une fois de plus que ce calcul est fondé sur les croyances du décideur politique, qui peuvent se révéler incorrectes. Je fais ces calculs en postulant différentes hypothèses sur la précision des résultats de l'évaluation d'impact. Cette valeur de l'information est utilisée pour déterminer quel programme serait sélectionné pour l'étude. Je calcule ensuite les *a posteriori* que le décideur politique acquerrait après avoir étudié ce programme, en supposant certains effets réels moyens pour chaque programme[6]. Enfin, j'estime la valeur de l'exploitation des *a priori* comme étant la différence de résultats des programmes finalement sélectionnés pour la mise en œuvre d'une évaluation *a posteriori* si l'on devait toujours choisir le bras de traitement qui a la plus grande valeur d'information par opposition au choix de la bonne intervention à étudier avec une probabilité de 50/50.

Le tabl. 1 résume les résultats pour plusieurs intervalles de confiance, *a priori* et valeurs moyennes réelles, « A » représentant l'impact moyen des transferts monétaires conditionnels dans cet exemple et « B » l'impact moyen des programmes de restauration scolaire. Les estimations sont exprimées en unités brutes – points de pourcentage d'augmentation des taux de scolarisation. Le tableau peut être lu comme suit : la première colonne indique l'*a priori* moyen supposé pour le programme A, l'*a priori* moyen pour le programme B étant supposé être nul ; les deuxième et troisième colonnes donnent des écarts-types hypothétiques pour les *a priori* des programmes A et B, respectivement ; la quatrième colonne fournit diverses valeurs hypothétiques pour l'impact réel du programme B, le programme A étant supposé avoir un impact nul ; les trois dernières colonnes présentent les avantages simulés de l'utilisation d'*a priori* pour déterminer le programme à évaluer en termes d'impact accru du programme sélectionné *a posteriori* pour plusieurs largeurs d'intervalle de confiance. La largeur des intervalles de confiance s'étend ici de haut en bas de la fourchette pour un niveau de confiance de 95 %. Par exemple, pour un intervalle de confiance de 0,1, l'erreur type serait égale à 0,1/(2*1,96). Je suppose que les décideurs sont bayésiens, de sorte que, lorsqu'ils déterminent le programme à mettre en œuvre après l'évaluation, ils le font en combinant de manière appropriée leurs *a priori* avec les nouvelles preuves fournies par l'évaluation d'impact.

6. Je limite l'attention à la moyenne, quoique d'autres parties de la distribution des résultats puissent également présenter un intérêt.

Tableau 1
Estimations des avantages de la prise en compte des a priori.

A priori				**Avantages de l'utilisation d'*a priori* pour différents intervalles de confiance**		
Moyenne pour A	**Écart-type pour A**	**Écart-type pour B**	**Moyenne B**	**0,01**	**0,1**	**1**
0,1	0,1	2,0	1	0,5	0,5	0,5
0,1	1,0	2,0	1	0,5	0,5	0,5
0,1	5,0	2,0	1	– 0,5	– 0,5	– 0,5
0,5	0,1	2,0	1	0,5	0,5	0,5
0,5	1,0	2,0	1	0,5	0,5	0,5
0,5	5,0	2,0	1	– 0,5	– 0,5	– 0,5
1,0	0,1	2,0	1	0,0	0,0	0,0
1,0	1,0	2,0	1	0,0	0,0	0,0
1,0	5,0	2,0	1	0,0	0,0	0,0
5,0	0,1	2,0	1	0,0	0,0	0,0
5,0	1,0	2,0	1	0,0	0,0	0,0
5,0	5,0	2,0	1	0,0	0,0	0,0
0,1	0,1	2,0	5	2,5	2,5	2,5
0,1	1,0	2,0	5	2,5	2,5	2,5
0,1	5,0	2,0	5	– 2,5	– 2,5	– 2,5
0,5	0,1	2,0	5	2,5	2,5	2,5
0,5	1,0	2,0	5	2,5	2,5	2,5
0,5	5,0	2,0	5	– 2,5	– 2,5	– 2,5
1,0	0,1	2,0	5	2,5	2,5	2,5
1,0	1,0	2,0	5	2,5	2,5	2,5
1,0	5,0	2,0	5	– 2,5	– 2,5	– 2,5
5,0	0,1	2,0	5	0,0	0,0	0,0
5,0	1,0	2,0	5	0,0	0,0	0,0
5,0	5,0	2,0	5	0,0	0,0	0,0
0,1	0,1	2,0	10	5,0	5,0	5,0
0,1	1,0	2,0	10	5,0	5,0	5,0
0,1	5,0	2,0	10	– 5,0	– 5,0	– 5,0
0,5	0,1	2,0	10	5,0	5,0	5,0
0,5	1,0	2,0	10	5,0	5,0	5,0
0,5	5,0	2,0	10	– 5,0	– 5,0	– 5,0
1,0	0,1	2,0	10	5,0	5,0	5,0
1,0	1,0	2,0	10	5,0	5,0	5,0
1,0	5,0	2,0	10	– 5,0	– 5,0	– 5,0
5,0	0,1	2,0	10	5,0	5,0	5,0
5,0	1,0	2,0	10	5,0	5,0	5,0
5,0	5,0	2,0	10	– 5,0	– 5,0	– 5,0

Source : Eva Vivalt.

Ces simulations suggèrent que les bénéfices peuvent être assez importants. Pour les valeurs que j'ai sélectionnées, les avantages étaient aussi importants que si le programme finalement choisi avait un effet supérieur de 5 points de pourcentage, c'est-à-dire qu'il augmentait les taux de scolarisation de 5 points de pourcentage, plutôt que de 0 point de pourcentage. Toutefois, les avantages sont largement subordonnés aux *a priori* et, bien sûr, aux effets moyens réels de chaque programme.

Dans cet exemple, l'exploitation des croyances préalables pourrait se traduire par un bénéfice attendu de la moitié de la différence entre les impacts réels des deux programmes potentiels. C'est logique : l'exploitation des *a priori* fournit aux individus une information qui les conduit à choisir l'un des deux programmes, alors que, sans exploitation des *a priori*, ils auraient choisi l'autre programme une fois sur deux. Cependant, l'exploitation des *a priori* n'est pas toujours utile. Pour certains *a priori*, le fait de recevoir des preuves empiriques tirées d'une évaluation d'impact ne suffirait pas à inciter les décideurs à prendre une meilleure décision *a posteriori*. S'ils ont plus d'incertitudes sur un programme particulier et qu'ils ne changent pas d'avis au vu des résultats de l'évaluation d'impact de ce programme (par exemple, si l'évaluation manquait de puissance statistique), l'utilisation d'*a priori* pour déterminer lequel évaluer pourrait tout de même amener un décideur à choisir le pire programme à mettre en œuvre après l'évaluation.

L'utilité de l'exploitation des *a priori* dépend donc en grande partie des *a priori* eux-mêmes. Le tabl. 2 présente des simulations utilisant les mêmes intrants, à l'exception de l'*a priori* moyen sur l'impact du programme A et de l'*a priori* moyen sur l'impact du programme B, chacun d'eux étant supérieur de 5 points de pourcentage à ceux des simulations présentées dans le tabl. 1. Cela étant, loin de se révéler utile pour la plupart des valeurs choisies, l'exploitation des *a priori* cause du tort la plupart du temps. Les valeurs choisies pour les distributions des *a priori* dans ce tableau pourraient refléter un excès d'optimisme quant aux résultats des programmes, qui a été observé dans plusieurs études (Vivalt et Coville, 2016 ; Casey *et al.*, 2018). Cependant, si ces biais étaient systématiques et prévisibles, ils pourraient être corrigés de façon à ne plus avoir d'effets délétères. Pour d'autres distributions d'*a priori*, l'exploitation des *a priori* n'est ni utile ni néfaste, mais n'a tout simplement aucun effet sur le processus décisionnel, comme dans les cas où un décideur a trop confiance en ses propres croyances. Là encore, à long terme, on pourrait s'attendre à ce que les décideurs soient subtils et capables de corriger au moins partiellement leur optimisme ou leur excès de confiance. Cela nécessiterait de nombreux changements dans les processus décisionnels, mais il y a déjà des signes indiquant que l'explicitation, la modélisation et l'utilisation d'*a priori* gagnent du terrain (par exemple, le projet Darpa Score : DellaVigna et Pope, 2018a ; 2018b ; DellaVigna *et al.*, 2019).

Tableau 2
Estimations des avantages pour différentes valeurs d'a priori.

A priori				**Avantages de l'utilisation d'*a priori* pour différents intervalles de confiance**		
Moyenne pour A	**Écart-type pour A**	**Écart-type pour B**	**Moyenne B**	**0,01**	**0,1**	**1**
5,1	0,1	2,0	1	– 0,5	– 0,5	– 0,5
5,1	1,0	2,0	1	– 0,5	– 0,5	– 0,5
5,1	5,0	2,0	1	0,5	0,5	0,5
5,5	0,1	2,0	1	– 0,5	– 0,5	– 0,5
5,5	1,0	2,0	1	– 0,5	– 0,5	– 0,5
5,5	5,0	2,0	1	0,5	0,5	0,5
6,0	0,1	2,0	1	– 0,5	– 0,5	0,0
6,0	1,0	2,0	1	– 0,5	– 0,5	– 0,5
6,0	5,0	2,0	1	0,5	0,5	0,5
10,0	0,1	2,0	1	– 0,5	– 0,5	0,0
10,0	1,0	2,0	1	– 0,5	– 0,5	– 0,5
10,0	5,0	2,0	1	0,5	0,5	0,5
5,1	0,1	2,0	5	– 2,5	– 2,5	– 2,5
5,1	1,0	2,0	5	– 2,5	– 2,5	– 2,5
5,1	5,0	2,0	5	2,5	2,5	2,5
5,5	0,1	2,0	5	– 2,5	– 2,5	– 2,5
5,5	1,0	2,0	5	– 2,5	– 2,5	– 2,5
5,5	5,0	2,0	5	2,5	2,5	2,5
6,0	0,1	2,0	5	– 2,5	– 2,5	0,0
6,0	1,0	2,0	5	– 2,5	– 2,5	– 2,5
6,0	5,0	2,0	5	2,5	2,5	2,5
10,0	0,1	2,0	5	– 2,5	– 2,5	0,0
10,0	1,0	2,0	5	– 2,5	– 2,5	– 2,5
10,0	5,0	2,0	5	2,5	2,5	2,5
5,1	0,1	2,0	10	0,0	0,0	0,0
5,1	1,0	2,0	10	0,0	0,0	0,0
5,1	5,0	2,0	10	0,0	0,0	0,0
5,5	0,1	2,0	10	0,0	0,0	0,0
5,5	1,0	2,0	10	0,0	0,0	0,0
5,5	5,0	2,0	10	0,0	0,0	0,0
6,0	0,1	2,0	10	0,0	0,0	5,0
6,0	1,0	2,0	10	0,0	0,0	0,0
6,0	5,0	2,0	10	0,0	0,0	0,0
10,0	0,1	2,0	10	– 5,0	– 5,0	0,0
10,0	1,0	2,0	10	– 5,0	– 5,0	– 5,0
10,0	5,0	2,0	10	5,0	5,0	5,0

Source : Eva Vivalt.

Conclusion

On pourrait souhaiter pratiquer un exercice similaire en utilisant des *a priori* réels sur diverses interventions et, de surcroît, pour étudier l'assignation déterministe. Bien que j'aie recueilli une variété de données sur des *a priori* dans le cadre d'autres projets (Vivalt et Coville, 2016 ; Coville et Vivalt, 2017), leur utilisation à cette fin n'est pas triviale, car, en général, les interventions visent à affecter différents indicateurs de résultats (par exemple, les taux de scolarisation par rapport à la prévalence des diarrhées) et, pour pouvoir faire des comparaisons entre différents indicateurs de résultats, il faut pouvoir évaluer l'importance relative des résultats, une tâche qui dépasse largement la portée de ce chapitre.

Je propose plutôt quelques remarques. Mon propos ici était de recenser les points communs entre les RCT et les non-RCT, même si j'ai également relevé des différences là où elles existent. Je ne me suis toutefois pas beaucoup concentrée sur ces différences, parce que je m'attends à ce que les avantages de l'assignation déterministe en matière de valeur d'information ne soient pas décisifs pour faire un choix. En particulier, il arrive qu'une RCT soit impossible à pratiquer et que seules des méthodes quasi expérimentales puissent être utilisées – dans ce cas, il n'y a pas véritablement de choix à faire. Parfois, c'est uniquement la volonté politique qui motive la RCT, de sorte que la question du choix est à nouveau discutable. En principe, je dirais qu'il vaut mieux envisager les méthodes au cas par cas, mais, dans la pratique, les chercheurs sont rarement en mesure de faire ce choix.

Alors que le débat oppose les RCT aux non-RCT, d'autres questions relatives à la conception des protocoles d'expérimentation paraissent potentiellement plus importantes et plus négligées. Premièrement, remarquablement peu de chercheurs en économie ou de décideurs politiques semblent être bayésiens, ce qui peut conduire à privilégier des résultats significatifs avec des effets de faible ampleur, par rapport à des programmes ayant des résultats moins certains, mais potentiellement plus importants. Deuxièmement, les RCT et les non-RCT sont généralement très limitées dans ce qu'elles étudient et elles excluent la plupart des effets indirects. Par exemple, supposons qu'un programme éducatif soit susceptible d'avoir une incidence sur les revenus ou sur la santé à l'âge adulte. Il est rare que quelqu'un revienne plusieurs années plus tard pour étudier ces effets. Il peut également y avoir des retombées sur les institutions ou les générations suivantes. La plupart du temps, ces possibilités sont ignorées. Il y a des raisons à cela : il est très difficile de cerner tous les effets pertinents. Néanmoins, il existe des méthodes qui pourraient être utilisées conjointement avec une RCT ou une non-RCT, comme l'approche du score de substitution d'Athey *et al.* (2016) pour estimer des résultats à long terme, qui sont encore négligés par les RCT comme par les non-RCT. Par ailleurs, les processus décisionnels sont loin d'être idéaux, et relativement peu de personnes prennent sérieusement en compte les preuves empiriques issues de RCT ou de non-RCT, sans parler des enseignements non biaisés qu'on peut en tirer. En outre, il reste encore beaucoup

à faire pour déterminer comment expliciter et agréger au mieux les prévisions afin de pouvoir les utiliser de façon bénéfique. Enfin, il est difficile de synthétiser les connaissances issues de plusieurs études tant que les chercheurs ne parviennent pas à se coordonner, par exemple quand ils ne font pas l'effort de partager les indicateurs de résultats communs. Moins de 10 % des études figurant dans la série de données d'AidGrade sur les résultats des évaluations d'impact en développement international ont publié leurs microdonnées sous-jacentes (Vivalt, 2020). Ironiquement, alors que les détracteurs des RCT accusent les partisans des RCT de ne pas avoir de vue d'ensemble, on pourrait très bien leur reprocher la même erreur. Il y a pourtant des batailles potentiellement plus importantes à mener.

Épilogue

La randomisation et l'évaluation des politiques sociales revisitées

James J. Heckman

Préambule

Ce chapitre actualise mon article publié en 1992, « Randomization and Social Policy Evaluation » (Heckman, 1992) et le replace dans le contexte de recherches ultérieures. Le papier demeure pertinent pour comprendre la nature fondamentale des expérimentations et les leçons à tirer des expérimentations « idéales » sans attrition, non-réponse et stratification sur les variables de *résultat* d'intérêt. Il vaut la peine d'être revisité à la lumière des controverses qui se sont fait jour autour du rôle de la randomisation en économie du développement. Les arguments conceptuels avancés ici n'ont pas été traités dans la littérature, à l'inverse de nombreux problèmes de mise en œuvre.

Ce préambule aborde à travers une perspective historique les expériences de terrain et les origines du mouvement expérimental en économie. L'histoire de l'expérimentation de terrain en économie depuis 1965 comprend deux grandes époques : (1) la première vague, qui s'est servie des expérimentations pour régler d'importants débats de politiques publiques lorsque les preuves non expérimentales étaient ambiguës et (2) la résurgence de l'expérimentation en économie du développement, qui a abouti au prix Nobel d'économie en 2019. Chaque époque a été marquée par un zèle quasi religieux en faveur de la méthodologie des évaluations par assignation aléatoire (*Randomized Controlled Trials* – RCT). C'est pourquoi je nomme ci-après ces deux époques « Grands Réveils », en l'honneur des deux renaissances religieuses qui ont façonné les Églises protestantes en Amérique du Nord aux XVIII^e^ et XIX^e^ siècles et en reconnaissance de la quête effrénée de pureté méthodologique constatée dans le domaine économique aux deux époques.

Le Premier Grand Réveil a eu lieu sous l'impulsion du mouvement pour l'évaluation des programmes d'emploi, d'éducation et de santé promulgués dans le cadre de la guerre contre la pauvreté de Lyndon Johnson. Le Second Grand Réveil a fait irruption en économie du développement dans le sillage d'une série de microprogrammes ciblant des pays moins développés et financés par des ONG, des milliardaires et diverses institutions internationales influentes. Les économistes qui promeuvent le Second Grand Réveil font peu de cas des enseignements tirés des limites des expérimentations sociales du Premier Grand Réveil. Les incitations et ambitions professionnelles de la nouvelle génération plaident contre l'examen et la citation des contributions et enseignements du Premier Réveil, lequel s'est terminé par une remise en question considérable de la revendication de « résultats transparents » et un déclin de l'enthousiasme irréfléchi suscité par les RCT. Le Second Réveil connaîtra probablement le même sort.

Le Premier Réveil

Bien avant que la randomisation ne devienne de rigueur dans le domaine du développement, au Premier Grand Réveil, elle était préconisée pour évaluer une grande diversité de programmes sociaux, d'interventions éducatives, de programmes de formation professionnelle et de réformes de l'aide sociale.

Au cours de la première vague, des cabinets d'étude de premier plan, tels que Westat, Mathematica, SRI, Abt Associates et Manpower Demonstration Research Corporation (MDRC), ont répondu à une demande de l'Office of Economic Opportunity (OEO), qui coordonnait la guerre contre la pauvreté de Lyndon Johnson, en évaluant une série de programmes sociaux nouvellement lancés. L'importance accordée à l'évaluation s'est propagée à de nombreuses agences fédérales américaines.

Ce premier effort d'évaluation a conduit à la collecte de nouveaux jeux de microdonnées de panel, qui continuent de guider la compréhension de la société et sont désormais largement imités dans le monde entier. Le Premier Réveil a également favorisé l'adoption de nouvelles méthodologies pour analyser les graves problèmes ayant entaché les expérimentations conduites lors de la première vague.

La première utilisation à grande échelle de la randomisation en économie a porté sur l'évaluation de programmes d'impôt négatif sur le revenu (INR). Ces programmes ont été proposés par Milton Friedman (FRIEDMAN, 2009, réédition) et d'autres comme alternative aux lourds programmes de transfert de l'aide sociale de l'époque, qui taxaient lourdement les travailleurs à faible revenu en réduisant considérablement les prestations sociales pour chaque dollar supplémentaire gagné. L'INR a été conçu pour remplacer le système morcelé d'aide sociale des années 1960 en accordant un transfert forfaitaire aux pauvres et en imposant les revenus supplémentaires à un taux bas et uniforme sur l'ensemble du barème des revenus. La question politique était de savoir si l'imposition de l'INR risquait de réduire sensiblement l'offre de main-d'œuvre. La réponse dépendait de la taille relative des effets de revenu et de substitution. Les transferts risquaient

de réduire l'offre de main-d'œuvre à travers un effet de revenu. En revanche, le taux d'imposition réduit sur les revenus risquait de l'accroître par un effet de substitution. Les estimations non expérimentales des effets de revenu et de substitution étaient à l'époque très disparates, comme le montre le chapitre d'introduction de CAIN et WATTS (1973).

Au début des années 1960, Heather Ross, alors étudiante de troisième cycle au MIT, a proposé une expérimentation randomisée à grande échelle pour mesurer les effets de l'INR. L'OEO a accepté sa proposition et l'a financée. De nombreux cabinets de conseil économique ont relevé le défi. La première expérimentation INR a vu le jour en 1968.

Les chercheurs des débuts se sont aventurés dans des eaux profondes et s'y sont parfois noyés. Les plans d'étude initiaux étaient défectueux. Les études étaient criblées de biais de sélection. Les taux d'attrition et de non-participation étaient massifs. Curieusement, l'analyse des données INR a contribué à lancer la discipline alors balbutiante de la micro-économétrie. Cette époque a culminé avec le témoignage de Cogan (Congress of the United States, SCOFSOPA, 1978) livré devant le Congrès américain, dans lequel il a réanalysé les données de l'expérience INR à l'aide des toutes nouvelles techniques micro-économétriques. Ces méthodes ont ensuite été consacrées par le comité Nobel en 2000.

Le témoignage de Cogan a remis en question les preuves « transparentes » de l'expérimentation en pointant une multitude de biais de sélection. Il a montré des impacts négatifs sur l'offre de main-d'œuvre qui étaient sensiblement plus importants que les impacts insignifiants constatés à partir des comparaisons expérimentales « transparentes » des différences moyennes entre les groupes de traitement et de contrôle. Lors de ces audiences, le sénateur Daniel Patrick Moynihan s'est dit consterné par la piètre qualité des preuves expérimentales « transparentes », révélées par l'analyse de Cogan, et lui a exprimé sa gratitude pour avoir présenté un rapport honnête de ce que l'expérimentation avait réellement démontré à l'appui de méthodes non expérimentales pour analyser les données expérimentales erronées.

Le Second Réveil

La seconde vague est aujourd'hui à son zénith. L'enthousiasme suscité par l'expérimentation a conduit les ONG, les fondations et les gouvernements à en prescrire l'application. Alors que la première vague était motivée par le désir de traiter des grandes questions sociales, la seconde vague a une orientation davantage méthodologique. Elle est le fruit de l'obsession professionnelle des économistes pour les « effets causaux », qu'ils recherchent coûte que coûte, même quand ceux qu'ils identifient sont dénués de signification sociale et/ou économique (introduction de Deaton, ce volume). Les études miniaturistes ont ainsi été couvertes de louanges, puisque jugées idéales pour une économie empirique rigoureuse. Il ne s'agissait plus de poser de grandes questions importantes et d'essayer d'y répondre, mais de chercher des réponses claires et nettes à des questions mineures sans grande conséquence politique. En effet, le comité

Nobel a salué en 2019 les praticiens de la seconde vague pour s'être concentrés sur des « problèmes plus petits et plus gérables » (Royal Academy of Sciences, 2019). Le prix Nobel est venu récompenser la pureté méthodologique et la « gérabilité » plutôt que la substance.

Il est utile ici de situer la quête de nombreux spécialistes de l'économie appliquée, qui paradent en tête du défilé de la seconde vague dans un cadre de régression traditionnel. Imaginons que Y soit un résultat d'intérêt. Supposons ensuite que :

$$Y = X\beta + D\alpha + U$$

où X est un vecteur de variables de contrôle observées, D est un indicateur de la réception d'un traitement ($D = 1$ si traité, $D = 0$ si non) et U est corrélé avec D. α est « l'effet » du traitement en contrôlant X et U. Si l'on ne tient pas compte de X et U, les estimations naïves de α sont biaisées et le signe du biais déterminé par le signe de la corrélation entre U et D en contrôlant X. La randomisation évite ce biais si elle est correctement menée.

Comme l'indique la littérature récente sur les variables instrumentales, dans le cadre du Second Réveil, l'élimination de ce biais est la préoccupation primordiale, généralement à l'exclusion de l'enjeu de savoir si α répond à une question importante, que ce soit en théorie ou en pratique. Dans le Premier Réveil, cet enjeu était au premier plan.

Le papier révisé présenté ici, tout comme un article de suivi de HECKMAN et SMITH (1995) ont été rédigés après la première vague d'enthousiasme en faveur des RCT et avant la seconde vague. Ces deux articles n'ont rien perdu de leur pertinence aujourd'hui. Le fait que la seconde vague ait vu le jour est un hommage, soit à la mauvaise rédaction de ces articles, soit à la capacité démontrée des économistes à ignorer les leçons durement gagnées du passé et aux fortes incitations carriéristes poussant à verser du vieux vin dans des bouteilles neuves et à en oublier l'origine. Je passe maintenant au papier original.

Introduction

Ce papier examine les avantages et les limites de l'expérimentation sociale *randomisée* en tant qu'outil d'évaluation de programmes sociaux[1]. L'argument en faveur de l'expérimentation sociale est désormais familier. Les données transversales et chronologiques disponibles présentent souvent une variabilité insuffisante au niveau des variables explicatives critiques pour permettre aux

1. Tout au long de ce papier, je m'abstiens de reprendre des arguments familiers sur les limites des expérimentations sociales et me concentre sur un problème qui n'est pas traité dans la littérature consacrée à ce sujet. Voir COOK et CAMPBELL (1979), les papiers dans HAUSMAN et WISE (1985a) et les autres chapitres de ce volume pour de plus amples informations sur les problèmes d'attrition, les effets d'entraînement, etc.

analystes de proposer des estimations convaincantes des impacts des programmes sociaux sur les variables de *résultat* cibles. En collectant des données en vue d'induire davantage de variations dans les variables explicatives, on obtient des estimations plus précises des impacts des politiques. En outre, la variation contrôlée des variables explicatives peut rendre exogènes des variables endogènes, c'est-à-dire qu'elle peut induire une variation indépendante des variables observées par rapport aux variables non observées. Les expérimentations sociales induisent une variation en contrôlant la manière dont les données sont collectées. La randomisation permet certes d'instiller une variation supplémentaire, mais ce n'est en aucun cas le seul moyen, ni même nécessairement le meilleur moyen d'obtenir la variation souhaitée.

Le point de départ de l'expérimentation sociale a été le paradigme de planification sociale de HAAVELMO (1944), MARSCHAK (1953) et TINBERGEN (1956). À l'époque, les connaissances en sciences sociales ont été jugées suffisamment avancées pour pouvoir identifier des relations comportementales de base qui, une fois estimées, pourraient être utilisées afin d'évaluer les impacts de toute une série de programmes sociaux, dont aucun n'avait été réellement mis en œuvre au moment de l'évaluation. L'approche par estimation d'« équations structurelles » appliquée à l'évaluation de politiques sociales promettait d'autoriser les analystes à simuler un large éventail de contrefactuels susceptibles de servir de base à des politiques sociales « optimales ». L'objectif de l'expérimentation sociale, tel que le voyaient CONLISK et WATTS (1969) et CONLISK (1973), était d'élaborer de meilleures estimations des équations structurelles requises pour effectuer la simulation des contrefactuels.

Les premiers partisans de la méthode expérimentale en économie se sont concentrés sur l'incapacité des études en coupes de l'offre de main-d'œuvre à isoler les effets « revenu » des effets de « substitution » nécessaires pour estimer l'impact des impôts négatifs sur le revenu (INR) sur l'offre de main-d'œuvre. Les expérimentations ont été conçues pour induire une plus grande variation des salaires et des revenus entre individus afin d'autoriser une meilleure estimation des paramètres de politiques critiques. L'objectif initial de ces expérimentations n'était pas d'évaluer un ensemble spécifique de programmes INR, mais d'estimer les paramètres qui pourraient être utilisés pour évaluer les impacts de ces programmes et de nombreux autres programmes possibles.

Au fur et à mesure que les expérimentations INR ont été mises en œuvre, leurs administrateurs ont commencé à en attendre moins. L'attention s'est peu à peu portée sur l'évaluation des effets de traitements spécifiques réellement en place (CAIN, 1975). Par extrapolation et interpolation, les effets de traitement estimés ont remplacé les simulations de politiques contrefactuelles fondées sur des paramètres structurels estimatifs et sont devenus la méthode de référence pour évaluer les programmes proposés non réellement mis en œuvre (HAUSMAN et WISE, 1985b).

Les arguments plus récents en faveur des expérimentations sociales randomisées représentent un recul spectaculaire par rapport à l'ambitieux programme

d'analyse des politiques sociales « optimales », qui n'a jamais été pleinement adopté par la plupart des économistes et qui a été totalement ignoré par d'autres spécialistes des sciences sociales. Un scepticisme considérable a récemment été exprimé quant à la valeur des méthodes économétriques ou statistiques pour estimer les impacts de programmes sociaux spécifiques ou les paramètres d'équations « structurelles » pour simulate des programmes sociaux qui n'ont pas encore vu le jour. Les études influentes de LALONDE (1986) et de FRAKER et MAYNARD (1987) en ont en effet convaincu plus d'un que les méthodes économétriques et statistiques sont incapables d'estimer les véritables impacts des programmes à partir de données non randomisées.

Les récents partisans des expérimentations sociales ont des ambitions plus modestes que les partisans des premiers jours. Ils proposent d'utiliser la randomisation pour évaluer des programmes réellement en place (qu'il s'agisse de programmes en cours ou de projets pilotes de « démonstration ») et d'éviter d'invoquer la litanie d'hypothèses souvent peu convaincantes qui sous-tendent les approches « structurelles », « économétriques » ou « statistiques » des évaluations de programmes[2]. Leur argument en faveur de la randomisation est extrêmement simple et convaincant : assigner des personnes au hasard à un programme et comparer les réponses cibles des participants à celles de non-participants exclus par randomisation. La différence moyenne entre les participants et les non-participants randomisés donne l'effet du programme. La recherche de paramètres « structurels profonds » est abandonnée. Nul besoin d'ajustements statistiques élaborés ou d'hypothèses arbitraires sur les formes fonctionnelles des équations pour estimer le paramètre d'intérêt avec des données randomisées. Et aucune stratégie d'estimation compliquée n'est nécessaire. Tout le monde comprend les moyennes. La randomisation garantit l'absence de biais de sélection parmi les participants, c'est-à-dire qu'il n'y a pas de sélection à l'entrée ou à la sortie du programme sur la base des résultats de l'échantillon randomisé.

Les partisans des expérimentations sociales randomisées font implicitement une hypothèse importante : que la randomisation ne modifie pas le programme étudié. Pour certains problèmes d'évaluation et pour certains modèles comportementaux, cette hypothèse est soit valable, soit inoffensive. Il en va différemment pour d'autres problèmes et modèles. L'une des principales conclusions de ce chapitre est que les adeptes de la randomisation ont surévalué leurs arguments d'avoir évité les hypothèses arbitraires. L'évaluation par randomisation formule des hypothèses comportementales implicites qui, dans certains contextes, sont assez fortes. Le biais induit par la randomisation est une possibilité réelle. Et tout indique qu'il s'agit d'un phénomène important.

En outre, les adeptes de la randomisation supposent implicitement que certaines différences moyennes dans les résultats sont invariablement les objets d'intérêt de l'évaluation. De fait, il existe de nombreux paramètres d'intérêt potentiel, dont certains seulement peuvent être intégrés à un cadre de différence moyenne.

2. Dans une ancienne contribution, ORCUTT et ORCUTT (1968) suggèrent cette utilisation des expérimentations sociales.

Néanmoins, les méthodes expérimentales *ne peuvent pas* estimer des différences médianes ou d'autres « effets de traitement par quantiles » sans invoquer des hypothèses plus fortes que celles qui sont nécessaires à l'obtention de moyennes. Les paramètres d'intérêt peuvent ne pas être définis par une randomisation hypothétique et les données randomisées peuvent ne pas être idéales pour estimer ces paramètres.

Les partisans de la randomisation sont souvent peu loquaces sur une question pratique importante. Bon nombre des programmes sociaux se déroulent en plusieurs étapes. À quel stade la randomisation doit-elle intervenir : lors de l'inscription, de l'assignation au traitement, de la promotion, de l'examen des performances ou du placement ? La réponse à cette question révèle une contradiction dans la justification des expérimentations randomisées. Pour utiliser des méthodes simples (c'est-à-dire des différences moyennes entre participants et non-participants) afin d'évaluer les effets des différentes étapes d'un programme, il faut procéder à une randomisation à chaque étape. Une telle randomisation en plusieurs étapes a rarement été mise en œuvre, probablement parce qu'elle modifierait radicalement le programme évalué[3]. Mais si une seule randomisation peut être pratiquée, l'évaluation de toutes les étapes d'un programme impose l'application de la méthodologie économétrique très controversée que l'on a récemment voulu éviter dans l'expérimentation sociale.

L'objectif de ce chapitre est d'expliciter les arguments pour et contre les expérimentations sociales randomisées. Afin de guider la discussion, je présente d'abord un prototype de programme social et examine les caractéristiques du programme qui intéressent les évaluateurs de politiques. Dans la deuxième section, j'aborde les difficultés qui se posent pour déterminer les caractéristiques d'intérêt du programme. Un énoncé précis du problème d'évaluation est dressé. Dans la section suivante, j'expose les arguments en faveur de la randomisation simple, puis j'examine les hypothèses comportementales implicites qui la sous-tendent et les conditions dans lesquelles elles se vérifient. Je discute également de ce qui peut et ne peut pas être appris d'une expérimentation sociale randomisée, même dans des conditions idéales. Dans la quatrième section, je présente quelques preuves empiriques indirectes sur la validité des hypothèses en prenant l'exemple d'une récente évaluation de la loi sur les partenariats en matière de formation professionnelle (*Job Training Partnership Act* – JTPA). J'examine également certaines études comparatives de leur validité menées sur des essais cliniques randomisés en médecine. Dans la cinquième section, j'aborde la question du choix du stade auquel il convient de procéder à la randomisation dans un programme en plusieurs étapes. Dans la sixième section, je traite de la tension entre les anciens et les nouveaux arguments en faveur de l'expérimentation sociale. La dernière section résume l'argumentaire présenté.

3. Voir cependant l'évaluation du programme ABC : Ramey *et al.* (1976), qui comporte une randomisation en plusieurs phases.

Questions d'intérêt dans l'évaluation d'un prototype de programme social

Le prototype considéré ici est un programme de formation de main-d'œuvre analogue au programme JTPA décrit par HECKMAN *et al.* (1998b). Ce programme prototypique proposait différentes options de formation aux stagiaires potentiels. Des compétences professionnelles spécifiques peuvent être acquises, ainsi que des compétences générales (telles que la lecture, l'écriture, le calcul). Une formation générale de rattrapage peut précéder la formation spécifique. Un placement professionnel peut être proposé dans le cadre d'un service distinct, indépendamment de toute acquisition de compétences ou à l'issue de la formation. Certains programmes de compétences spécifiques impliquent de travailler pour un employeur à un salaire subventionné (soit une formation continue).

Les personnes qui suivent la formation passent par les étapes suivantes : elles postulent (1), sont acceptées (2), sont affectées à une séquence de formation spécifique (3), sont évaluées (4), sont certifiées dans une compétence (5) et sont placées chez l'employeur (6). Pour les stagiaires qui suivent une formation continue, les étapes 3 à 6 sont combinées, bien que les stagiaires puissent être périodiquement évalués pendant leur période de formation. Les stagiaires peuvent abandonner le processus ou en être exclus à chaque étape.

Les centres de formation étaient payés par le gouvernement américain sur la base de la qualité du placement de leurs stagiaires. La qualité était mesurée en partie par les salaires perçus pendant une période déterminée après l'achèvement du programme de formation (par exemple, six mois). Les responsables étaient ainsi incités à former des personnes ayant des chances de trouver un emploi de qualité et pouvant y parvenir à faible coût pour le centre. Les stagiaires ont perçu une rémunération (subventions) pendant la durée du programme. Les centres de formation ont recruté des stagiaires par le biais de divers mécanismes de promotion.

Il y a de nombreuses questions d'intérêt pour les évaluateurs du programme. Celle qui reçoit le plus d'attention est l'effet de la formation sur les personnes formées.

Q-1 Quel est l'effet de la formation sur les personnes formées ?

C'est l'objectif central mis en avant dans de nombreuses évaluations. Lorsque les coûts d'un programme sont soustraits de la réponse à la question Q-1 et que les rendements sont correctement escomptés, on obtient le bénéfice net du programme pour un groupe de stagiaires donné.

Mais il y a beaucoup d'autres questions qui revêtent également un intérêt potentiel pour les évaluateurs du programme, par exemple :

Q-2 Quel est l'effet de la formation sur les stagiaires aléatoirement assignés ?

La réponse à la question Q-2 serait d'un grand intérêt si la formation était obligatoire pour toute une population, comme dans les programmes de travail obligatoire qui forcent les bénéficiaires de l'aide sociale à suivre une formation. D'autres questions d'intérêt concernent les décisions de candidature.

Q-3 Quel est l'effet des subventions (et/ou de la publicité, des conditions du marché du travail local, du revenu familial et/ou de la race, du sexe) sur les décisions de candidature ?

Q-4 Quels sont les effets des normes de performance du centre, des taux de profit, de la structure du marché du travail local et du contrôle gouvernemental sur l'acceptation par les centres de formation des décisions des candidats et sur leur placement dans des programmes spécifiques ?

Q-5 Quels sont les effets du contexte familial, des taux de profit du centre, des subventions et des conditions du marché du travail local sur la décision d'abandonner un programme et sur le temps nécessaire pour terminer le programme ?

Q-6 Quels sont les effets des conditions du marché du travail, des subventions, des taux de profit, etc. sur les taux de placement et les niveaux de salaire et de temps de travail atteints lors du placement ?

Q-7 Quel est le coût de la formation d'un travailleur selon les différentes modalités possibles ?

Les réponses à toutes ces questions et leur affinement sont susceptibles d'intéresser les décideurs politiques. Le problème central de l'évaluation est de déterminer comment obtenir des réponses convaincantes à ces questions.

Le problème de l'évaluation

Pour définir les caractéristiques essentielles du problème d'évaluation, il est préférable de se concentrer uniquement sur quelques-unes des questions énumérées ci-dessus. Je porte donc mon attention sur les questions Q-1 et Q-2 et sur une combinaison des ingrédients des questions Q-3 et Q-4.

Q-3' Quels sont les effets des variables énumérées en Q-3 et Q-4 sur la candidature et l'inscription des personnes ?

Pour simplifier l'analyse, je suppose tout au long de la discussion de cette section qu'il n'existe qu'un seul type de traitement administré par le programme, de sorte que la détermination de l'assignation au traitement ne pose pas de problème. Je suppose qu'il n'y a pas d'attrition dans le programme et que la durée de la participation au programme est fixe. Ces hypothèses seraient vraies si, par exemple, le programme idéal se déroulait à un seul instant T dans le temps et donnait à chaque participant la même « dose », bien que la réponse à la dose puisse varier d'une personne à l'autre. Je suppose également l'absence

de toute interdépendance entre les unités résultant d'inobservables ou d'effets de rétroaction, communs ou propres à chaque site[4].

Ce papier ne s'intéresse pas exclusivement, ni même principalement, à l'« estimation structurelle », parce qu'elle n'est pas préconisée dans la littérature récente sur les expérimentations sociales et parce qu'une discussion sur ce sujet soulève des questions supplémentaires qui ne sont pas pertinentes ici. Les approches structurelles exigent la spécification d'un ensemble commun de caractéristiques et d'un modèle de participation aux programmes et de résultats pour décrire tous les programmes d'intérêt potentiel. Elles requièrent d'estimer les réponses aux variations de caractéristiques qui décrivent des programmes non encore déployés, ce qui nécessite à son tour la spécification et la mesure d'un ensemble commun de caractéristiques sous-tendant ces programmes.

L'approche structurelle prototypique est bien illustrée dans les premiers travaux sur l'estimation des réponses de l'offre de main-d'œuvre aux programmes d'impôt négatif sur le revenu. Ces programmes fonctionnaient en modifiant le niveau de salaire et de revenu des participants potentiels. Selon la théorie néoclassique de l'offre de main-d'œuvre, si l'on peut déterminer la réponse de l'offre de main-d'œuvre aux changements de salaires et de niveaux de revenus (les effets de « substitution » et de « revenu », respectivement), on peut également déterminer qui participerait à un programme (voir, par exemple, ASHENFELTER, 1983). Ainsi, à partir d'un ensemble commun de paramètres, on peut simuler l'effet de *tous* les programmes INR *possibles* sur l'offre de main-d'œuvre.

C'est pour cette raison que les premiers partisans des expérimentations sociales ont cherché à concevoir des expérimentations qui produiraient des variations maximales (indépendantes des échantillons) des niveaux de salaire et de revenu entre les sujets, de manière à obtenir des estimations précises des effets sur les salaires et les revenus. CAIN et WATTS (1973) ont fait valoir que, dans les données en coupes, la variation des salaires et des revenus était suffisamment faible pour qu'il soit difficile, voire impossible, d'estimer des effets distincts des salaires et des revenus sur l'offre de main-d'œuvre.

L'approche structurelle est très séduisante lorsqu'elle est crédible. Elle se focalise sur des aspects essentiels de la réponse aux programmes, mais son utilisation pratique nécessite la formulation d'hypothèses comportementales fortes afin de placer différents programmes sur une base commune. En outre, elle exige que les caractéristiques communes des programmes puissent être mesurées. Tant les problèmes que les hypothèses comportementales requises dans l'approche structurelle soulèvent des questions qui dépassent le cadre de ce chapitre. Je me cantonne essentiellement au problème pratique – et néanmoins très difficile – de l'évaluation des effets de programmes existants et des réponses aux changements de paramètres de ces programmes qui pourraient affecter la participation aux programmes.

4. Il s'agit de l'hypothèse *Stable Unit Treatment Value Assumption* (SUTVA) de Rubin (HOLLAND, 1986). Elle est largement invoquée dans la littérature en économétrie et en statistique même si, de toute évidence, elle se révèle souvent fausse (HECKMAN *et al.*, 1998a).

Un modèle d'évaluation de programme

Pour être plus précis, il convient de définir la variable $D = 1$ si une personne participe à un programme hypothétique ; sinon $D = 0$. Si une personne participe, elle reçoit le résultat Y_1 ; sinon, elle reçoit Y_0. Ainsi, le résultat observé Y est :

$$\begin{aligned} Y &= Y_1 \quad \text{si} \quad D = 1 \\ Y &= Y_0 \quad \text{si} \quad D = 0 \end{aligned} \tag{1}$$

Une caractéristique essentielle du problème de l'évaluation est que nous n'observons pas la même personne dans les deux situations. C'est ce que certains statisticiens appellent le « problème de la référence causale » (voir, par exemple, HOLLAND, 1986). Faisons en sorte que Y_1 et Y_0 soient respectivement déterminés par X_1 et X_0. On peut supposer que X_1 induit des aspects pertinents de la formation reçue par les stagiaires. X_0 et X_1 peuvent contenir des variables en termes d'antécédents et de marché du travail local. Nous écrivons des fonctions reliant ces variables à Y_0 et Y_1, respectivement :

$$Y_1 = g_1(X_1), \tag{2a}$$

$$Y_0 = g_0(X_0). \tag{2b}$$

Pour les équations linéaires plus familières, (2a) et (2b) peuvent être spécialisées respectivement en :

$$Y_1 = X_1\beta_1 \tag{2a'}$$

et

$$Y_0 = X_0\beta_0 \tag{2b'}$$

Supposons que Z désigne des variables déterminant la participation au programme. Si

$$Z \in \Psi, D = 1 \; ; \; Z \notin \Psi, D = 0, \tag{3}$$

où Ψ est un ensemble de valeurs Z possibles. Si les personnes ont des caractéristiques qui se situent dans l'ensemble Ψ, elles participent au programme ; sinon, elles n'y participent pas. Z comprend les caractéristiques des personnes et leurs débouchés sur le marché du travail, ainsi que les caractéristiques des sites de formation sélectionnant les candidats. Afin d'économiser sur les symboles, je représente l'ensemble des variables explicatives par $C = (X_0, X_1, Z)$. Si une variable dans C n'apparaît pas dans X_1 ou X_0, son coefficient ou sa dérivée associée dans g_1 ou g_0 est fixé(e) à zéro pour toutes les valeurs de la variable.

Même si l'on pouvait observer toutes les composantes de C pour chaque personne d'un échantillon, on ne pourrait peut-être pas déterminer g_1, g_0 et Ψ. Les échantillons disponibles pourraient ne pas contenir suffisamment de variations dans les composantes de ces vecteurs pour retrouver g_0, g_1 ou pour identifier l'ensemble Ψ. C'est un problème de « multicolinéarité » (dans les variables de revenu et de salaire nécessaires pour déterminer les équations de l'offre de main-d'œuvre) et un manque de variation des revenus au sein de l'échantillon qui ont en partie motivé les partisans initiaux des expérimentations sociales en économie.

En supposant une variabilité suffisante dans les composantes des variables explicatives, on peut utiliser les données sur les participants pour déterminer g_1, sur les non-participants pour déterminer g_0 et l'échantillon combiné pour déterminer Ψ. En connaissant ces fonctions et ensembles, on peut facilement répondre aux problèmes d'évaluation Q-1, Q-2, et Q-3' (à condition que le support des variables X_1, X_0, et Z dans l'échantillon couvre le support de ces variables dans les populations cibles d'intérêt). Il serait ainsi possible d'obtenir Y_1 et Y_0 pour chaque personne et d'estimer le gain brut de participation pour chaque participant ou pour chaque personne de l'échantillon. De cette manière, les questions Q-1 et Q-2 peuvent trouver une réponse complète. De même, en connaissant Ψ, il est possible de répondre complètement à la question Q-3' pour chaque personne.

En pratique, les analystes n'observent pas toutes les composantes de C. Les composantes non observées de ces résultats et des fonctions de participation constituent des sources majeures de problèmes d'évaluation. Ce sont ces composantes manquantes qui motivent le traitement de Y_1, Y_0 et D comme des variables aléatoires, en fonction des informations disponibles. Ce caractère intrinsèquement aléatoire exclut toute stratégie visant à déterminer Y_1 et Y_0 pour chaque personne. Au lieu de cela, une approche statistique est adoptée pour estimer la distribution conjointe de Y_1, Y_0, D en fonction des informations disponibles ou de certaines de leurs caractéristiques.

Prenons a comme notation des informations disponibles. Ainsi, C_a contient les variables dont l'analyste *dispose* et qu'il juge légitimes pour déterminer Y_1, Y_0 et D. Ces variables peuvent être constituées de certaines composantes de C ainsi que de variables de substitution pour les composantes manquantes.

La distribution conjointe de Y_1, Y_0, D, compte tenu de $C_a = c_a$, est

$$F\left(y_0, y_1, d \middle| c_a\right) = \Pr\left(Y_0 \leqslant y_0, Y_1 \leqslant y_1, D = d \middle| C_a = c_a\right), \tag{4}$$

où je suis la convention en désignant les variables aléatoires par des lettres majuscules et leur réalisation par des lettres minuscules. Si (4) peut être déterminé, et que la distribution de C_a est connue, il est possible de répondre aux questions Q-1, Q-2, et Q-3' dans le sens suivant : on peut déterminer la distribution en population de Y_0, Y_1 et la *distribution* en population du gain brut de la participation au programme

$$\Delta = Y_1 - Y_0,$$

et on peut écrire la probabilité de l'événement $D = d$ compte tenu de Z_a.

Paramètres d'intérêt dans l'évaluation de programmes

Nous pouvons répondre à la question Q-1 si nous pouvons identifier

$$F\left(y_0, y_1 \middle| D = 1, c_a\right),$$

et donc

$$F\left(\delta \middle| D = 1, c_a\right)$$

(la distribution de l'effet du traitement sur les personnes traitées, où δ est la version minuscule de Δ). On peut répondre à la question Q-2 si nous connaissons

$$F\left(y_0, y_1 | c_a\right), \tag{5}$$

qui peut être produite à partir de (4) et la distribution des variables explicatives par des opérations de probabilité élémentaires. En ce sens, on peut déterminer les gains d'un déplacement aléatoire d'une personne depuis une distribution $F\left(y_0|c_a\right)$ vers une autre $F\left(y_1|c_a\right)$. La réponse à Q-3' peut être obtenue en calculant à partir de (4) la probabilité de participation :

$$\Pr\left(D=1|c_a\right)=F\left(d|c_a\right).$$

En pratique, les comparaisons de moyennes accaparent la majeure partie de l'attention dans la littérature, bien que les médianes, ou d'autres quantiles, soient également intéressantes. Une grande partie de la littérature *définit* la réponse à Q-1 comme :

$$E\left(\Delta|D=1, c_a\right)=E\left(Y_1-Y_0|D=1, c_a\right) \tag{6}$$

et la réponse à Q-2 comme

$$E\left(\Delta|c_a\right)=E\left(Y_1-Y_0|c_a\right), \tag{7}$$

quoique, en principe, il puisse être souhaitable de connaître la distribution complète de Δ, ou de certaines caractéristiques autres que la moyenne (par exemple, la médiane).

Même si les moyennes en (6) et (7) étaient nulles, il serait intéressant de savoir quelle fraction des participants ou de la population bénéficierait d'un programme. Pour cela, il faudrait connaître $F\left(\delta|D=1, c_a\right)$ ou $F\left(\delta|c_a\right)$, respectivement. Afin de vérifier la présence d'« écrémage » (phénomène par lequel les sites de formation sélectionnent les meilleurs candidats à un programme – ceux qui ont des valeurs élevées de Y_0 et Y_1), il est nécessaire de connaître la corrélation ou la dépendance stochastique entre Y_1 et Y_0. Il faudrait pour ce faire connaître les caractéristiques de :

$$F\left(y_1, y_0|D=1, c_a\right)$$

ou

$$F\left(y_1, y_0|c_a\right),$$

autres que les moyennes de Y_1 et Y_0. Pour répondre à de nombreuses questions, la connaissance des différences moyennes est insuffisante ou incomplète.

La détermination de la distribution conjointe (4) est un problème difficile. Dans la section suivante, je montre que les expérimentations sociales randomisées du type de celles présentées dans la littérature récente ne produisent pas de données suffisantes pour résoudre ce problème.

Les données produites de façon routinière à partir des registres de programmes sociaux permettent aux analystes de déterminer :

$$F\left(y_1|D=1, c_a\right),$$

la distribution des résultats pour les participants et

$$F\left(y_0 | D = 0, c_a\right),$$

la répartition des résultats pour les non-participants et ceux-ci sont parfois suffisamment riches pour déterminer

$$\Pr\left(D = 1 | c_a\right) = F\left(d | c_a\right),$$

la probabilité de participation. Mais, à moins que des informations supplémentaires ne soient disponibles, ces éléments ne suffisent pas à déterminer (4). En vertu de (1), il n'existe pas de données sur les deux composantes de $\left(y_1, Y_0\right)$ pour la même personne. En général, pour les mêmes valeurs de $C_a = c_a$

$$F\left(y_0 | D = 1, c_a\right) \neq F\left(y_0 | D = 0, c_a\right) \tag{8a}$$

et

$$F\left(y_1 | D = 1, c_a\right) \neq F\left(y_1 | D = 0, c_a\right), \tag{8b}$$

ce qui pose le problème du biais de sélection dans les distributions des résultats. Le problème de sélection se pose le plus souvent en termes de moyennes :

$$E\left(\Delta | D = 1, c_a\right) \neq E\left(Y_1 | D = 1, c_a\right) - E\left(Y_0 | D = 0, c_a\right) \tag{9a}$$

$$E\left(\Delta | c_a\right) \neq E\left(Y_1 | c_a\right) - E\left(Y_0 | c_a\right), \tag{9b}$$

c'est-à-dire que les personnes qui participent à un programme sont différentes des personnes qui n'y participent pas, en ce sens que les résultats moyens des participants dans la situation de non-participation seraient différents de ceux de non-participants, même après ajustement de C_a.

De nombreuses méthodes ont été proposées pour résoudre le problème de la sélection, que ce soit pour des moyennes ou pour des distributions entières. HECKMAN et ROBB (1985 ; 1986), HECKMAN et HONORÉ (1990), HECKMAN (1990a ; 1990b) et HECKMAN *et al.* (1997a) proposent des traitements alternatifs et complets des différentes approches de ce problème en économétrie et en statistique. Certaines hypothèses *a priori* non vérifiables doivent être formulées pour récupérer les composantes manquantes de la distribution. La construction de ces contrefactuels est inévitablement source de controverses.

Pour LALONDE (1986) et FRAKER et MAYNARD (1987), ces controverses ne se cantonnent pas à la sphère académique. Dans des travaux influents analysant des données expérimentales randomisées à l'aide de méthodes non expérimentales, ces auteurs produisent un large éventail d'estimations des impacts d'un même programme à l'appui de différentes méthodes non expérimentales. Selon eux, il n'est pas possible de faire un choix parmi des estimateurs non expérimentaux concurrents.

HECKMAN et HOTZ (1989) réanalysent leurs données et démontrent que leurs propos sont grandement exagérés. Aucun des deux collectifs d'auteurs n'a effectué de tests de spécification de modèle standard pour leurs estimations alternatives non expérimentales. Lorsque de tels tests sont effectués, ils permettent une estimation de tous les modèles, à l'exception des modèles non expérimentaux qui reproduisent l'inférence obtenue par des méthodes expérimentales.

Il y a néanmoins un fond de vérité dans les critiques de LALONDE (1986) et de FRAKER et MAYNARD (1987). Chaque test de modèle non expérimental proposé par HECKMAN et HOTZ (1989) a ses limites. Le test des caractéristiques de suridentification d'un modèle peut perdre de sa valeur dès lors que l'on change le modèle pour une forme exactement identifiée, une critique qui vaut également dans l'application du test de Durbin-Wu-Hausman (DURBIN, 1954 ; WU, 1973 ; HAUSMAN, 1978).

Toutes les méthodes non expérimentales reposent sur une hypothèse retenue et non vérifiable. La principale source d'intérêt des expérimentations randomisées est qu'elles *semblent* ne nécessiter aucune hypothèse. Dans la section suivante, je démontre que les arguments en faveur des évaluations randomisées reposent sur des hypothèses non déclarées concernant le problème d'intérêt, le nombre d'étapes d'un programme et les réactions des agents à la randomisation. Ces hypothèses sont différentes des hypothèses avancées dans la littérature non expérimentale en économétrie et en statistique, mais pas nécessairement plus crédibles.

Arguments pour et contre les expérimentations sociales randomisées

Les arguments en faveur des expérimentations sociales randomisées sont presque toujours énoncés dès lors que l'on tente de répondre aux questions Q-1 et Q-2 – le « problème causal » tel que défini par les statisticiens (FISHER, 1935 ; COX, 1958 ; RUBIN, 1978 ; HOLLAND, 1986). De ce point de vue, l'équation de participation qui répond à Q-3' est une « fonction de nuisance » qui peut générer un problème de sélection. La randomisation simple rend le statut du traitement statistiquement indépendant de (Y_1, Y_0, C).

Pour exposer le plus clairement possible les arguments en faveur de la randomisation, il est utile d'introduire une variable *A* indiquant la participation réelle à un programme :

$A = 1$ si une personne participe
$= 0$ sinon

et de la différencier de la variable *D* indiquant qui aurait participé à un programme obéissant à un protocole non expérimental. Supposons que *D** désigne une variable indiquant si un agent est susceptible d'être randomisé (c'est-à-dire si l'agent a postulé et a été accepté dans un régime de sélection aléatoire) :

*D** = 1 si une personne est susceptible d'etre randomisé
= 0 sinon.

Dans l'approche standard, la randomisation est mise en œuvre à un stade où *D** est révélé. Étant donné que $D^* = 1$, *A* est supposé être indépendant de (Y_0, Y_1, C), alors

$$F(y_0, y_1, c, a | D^* = 1) = F(y_0, y_1, c | D^* = 1) F(a | D^* = 1).$$

Des protocoles de randomisation plus élaborés pourraient être mis en œuvre, mais sont rarement proposés.

Le fait de modifier le processus d'inscription au programme en refusant de manière aléatoire l'accès aux personnes qui postulent et sont jugées aptes à participer à un programme peut rendre la distribution de D^* différente de celle de D. Une telle randomisation modifie l'ensemble des informations dont disposent les candidats potentiels et des administrateurs de programmes, à moins que ni les uns ni les autres ne soient informés de la possibilité d'une randomisation – une condition peu probable pour un programme en cours ou pour des programmes ponctuels dans de nombreux pays comme les États-Unis, où la loi impose la divulgation complète des règles de fonctionnement des programmes. Même s'il était possible d'agir à l'insu des stagiaires potentiels, il ne serait pas possible d'agir à l'insu les centres de formation qui administrent le programme. Rappelons que D^* est le résultat de décisions conjointes de stagiaires potentiels et de centres de formation. L'ensemble de conditionnement déterminant D^* diffère de celui de D par l'inclusion de la probabilité de sélection $\left(p = \Pr(A = 1)\right)$, c'est-à-dire qu'il inclut l'effet de la randomisation sur les choix des agents et des centres.

Les partisans de la randomisation invoquent l'hypothèse selon laquelle

$$\Pr\left(D = 1|c\right) = \Pr\left(D^* = 1|c,\, p\right), \qquad \text{(HY-1)}$$

ou supposent que c'est « pratiquement » vrai[5].

Il y a de nombreuses raisons de douter de la validité de cette hypothèse. Si des personnes qui auraient pu s'inscrire à un protocole non randomisé font des projets anticipant leur inscription à une formation, le fait d'ajouter de l'incertitude au stade de l'acceptation peut modifier leur décision de s'inscrire ou d'entreprendre des activités complémentaires à la formation. Les personnes peu enclines à prendre des risques auront tendance à être éliminées du programme. Même si la randomisation accroît l'utilité de l'agent[6], le comportement sera modifié. Si les centres de formation doivent randomiser après un processus de sélection, ils pourraient avoir à sélectionner davantage de personnes afin d'atteindre leurs objectifs de performance, ce qui pourrait amoindrir la qualité des stagiaires. Il peut y avoir dégradation de la qualité des candidats même si les places d'un programme sont rationnées. La randomisation peut résoudre les problèmes de rationnement de manière équitable s'il existe une file d'attente pour accéder au programme, mais elle peut aussi modifier la composition du vivier de stagiaires.

L'hypothèse (HY-1) est tout à fait naturelle dans le contexte de l'expérimentation agricole et biologique, dans lequel le modèle Fisher des expérimentations randomisées a vu le jour. Cependant, le modèle Fisher est un paradigme potentiellement trompeur en sciences sociales. Les humains agissent de manière intentionnelle

5. L'échec de cette hypothèse est un exemple de MARSCHAK (1953). Voir aussi la critique de LUCAS (1976) appliquée à l'expérimentation sociale. Il s'agit également d'un exemple d'effet « Hawthorne » (COOK et CAMPBELL, 1979).

6. Ce peut être le cas même si les agents sont adverses au risque en convexant un problème non convexe. Voir ARNOTT et STIGLITZ (1988).

et leur comportement est susceptible d'être modifié par l'introduction de la randomisation dans leur domaine de choix. Le modèle Fisher peut se révéler idéal pour l'étude des traitements à l'engrais sur les rendements des cultures. Les parcelles de terrain ne répondent pas à des traitements à l'engrais anticipés et ne peuvent pas non plus refuser d'être traitées. Les fabricants commerciaux d'engrais peuvent être exclus de la sélection de parcelles de terrain favorables dans un cadre expérimental agricole, tout comme les responsables de centres de formation ne peuvent être exclus de la sélection de stagiaires favorables dans un cadre de sciences sociales.

Si l'hypothèse (HY-1) est vraie,

$$F\left(y_1, c\middle|A=1\right) = F\left(y_1, c\middle|D^*=1\right) = F\left(y_1, c\middle|D=1\right), \tag{10a}$$

$$F\left(y_0, c\middle|A=0\right) = F\left(y_0, c\middle|D^*=1\right) = F\left(y_0, c\middle|D=1\right), \tag{10b}$$

$$E\left(Y_1\middle|A=1\right) - E\left(Y_0\middle|A=0\right) = E\left(\Delta\middle|D=1\right). \tag{11}$$

Les estimateurs simples de différence moyenne entre les participants et les non-participants randomisés répondent à la question Q-1 énoncée en termes de moyennes, du moins pour les grands échantillons. La distribution de variables explicatives *C* est la même dans les échantillons conditionnés en *A*. Les échantillons conditionnés en $A = 1$ et $A = 0$ sont ainsi équilibrés.

En ce sens, les données randomisées sont « idéales ». Les personnes non formées aux statistiques – comme les politiciens et les administrateurs de programmes – comprennent les moyennes et aucun ajustement statistique élaboré ni aucune hypothèse de forme fonctionnelle concernant un modèle ne sont imposés aux données. De plus, (11) *peut* être vraie, même si (HY-1) est fausse.

Cela se vérifie dans le modèle de variable endogène binaire largement employé (Heckman, 1978). En l'occurrence,

$$Y_1 = \alpha + Y_0. \tag{12}$$

Ce modèle est appelé « modèle à effets de traitement fixes pour toutes les unités » dans la littérature statistique (Cox, 1958). Ce modèle s'écrit :

$$Y_1 = g_1\left(x_1\right) = \alpha + g_0\left(x_0\right) = \alpha + Y_0,$$

de sorte que l'effet du traitement est le même pour tous. En ce qui concerne le modèle de régression linéaire de (2a') et (2b'), ce modèle peut s'écrire comme suit : $X_1\beta_1 = \alpha + X_0\beta_0$. Même si (HY-1) est fausse, (11) est vraie parce que

$$\begin{aligned}
&E\left(Y_1\middle|A=1\right) - E\left(Y_0\middle|A=0\right)\\
&= E\left(\alpha + Y_0\middle|A=1\right) - E\left(Y_0\middle|A=0\right)\\
&= \alpha + E\left(Y_0\middle|D^*=1\right) - E\left(Y_0\middle|D^*=1\right)\\
&= \alpha\\
&= E\left(\Delta\middle|D=1\right)\\
&= E\left(\Delta\right).
\end{aligned}$$

Le modèle à variables endogènes binaires est largement employé dans les travaux appliqués. La confiance placée en ce modèle renforce l'argument en vogue en faveur de la randomisation. Q-1 et Q-2 ont la même réponse dans ce modèle et la randomisation permet de répondre de manière convaincante aux deux questions.

L'exigence d'homogénéité des résultats du traitement peut être affaiblie et (11) peut encore se justifier si (HY-1) est fausse. Supposons qu'il existe un modèle de réponse aléatoire (parfois appelé modèle à effets aléatoires) :

$$Y_1 = Y_0 + (\alpha + \Xi), \tag{13a}$$

où Ξ est la réponse idiosyncrasique d'un individu au traitement après avoir éliminé une réponse commune α et

$$E(\Xi | D) = 0, \tag{13b}$$

alors (11) reste vrai. Si les stagiaires potentiels et les centres de formation ne connaissent pas les avantages du programme avant leur inscription au programme et qu'ils utilisent a $\alpha + \Xi$ pour prendre leur décision en matière de participation, alors (11) est toujours correcte. Ainsi, même si les réponses aux traitements sont hétérogènes, l'estimateur simple de différence moyenne obtenu à partir de données expérimentales peut toujours répondre à la version de différence moyenne de Q-1.

Il est important de noter les limitations des données obtenues à partir d'une expérimentation sociale « idéale », c'est-à-dire une expérimentation sociale vérifiant (HY-1). Sans invoquer d'autres hypothèses, on ne peut pas estimer la distribution de Δ conditionnelle ou inconditionnelle sur $D = 1$. On ne peut pas estimer la médiane de Δ ni déterminer l'importance empirique de l'« écrémage » (la dépendance stochastique entre Y_0 et Y_1) à partir des données, à moins de formuler l'hypothèse extrême d'une invariance de rang, c'est-à-dire un rang égal des personnes dans la distribution de Y_1 et dans la distribution de Y_0 (HECKMAN *et al.*, 1997a). Les données expérimentales et non expérimentales sont toujours confrontées au problème fondamental de l'impossibilité d'observer Y_0 et Y_1 pour une même personne. Les données expérimentales randomisées du type proposé dans la littérature ne facilitent que l'estimation simple d'un paramètre,

$$E(\Delta | D = 1, c).$$

Des hypothèses doivent être imposées pour produire des paramètres d'intérêt supplémentaires, même à partir de données expérimentales idéales. La réponse à la plupart des questions énumérées dans la première section requiert encore l'application de procédures économétriques, avec les hypothèses controversées qu'elles impliquent.

Si l'hypothèse (HY-1) n'est pas satisfaite, les égalités finales en (10a) et (10b) ne sont pas satisfaites et, en général

$$E(Y_1 | A = 1) - E(Y_0 | A = 0) \neq E(\Delta | D = 1).$$

De plus, les données produites par l'expérimentation ne permettront pas aux analystes d'évaluer les déterminants de la participation à un protocole non

randomisé, car les processus décisionnels de candidature et d'inscription auront été altérés par la randomisation, c'est-à-dire :

$$\Pr(D = 1|c) \neq \Pr(D^* = 1|c, p),$$

à moins que $p = 1$. Ainsi, l'expérimentation ne générera pas de données permettant de répondre à la question Q-3', à moins que la randomisation ne soit une caractéristique permanente du programme évalué.

Dans le cas général où la réponse des agents aux programmes est hétérogène $(\Xi \neq 0)$ et où les agents anticipent cette hétérogénéité (plus précisément, Ξ n'est pas stochastiquement indépendant de D), l'hypothèse (HY-1) joue un rôle crucial dans la justification des expérimentations sociales randomisées. Si le modèle (HY-1) est largement admis dans certains domaines scientifiques – comme l'expérimentation agricole, qui a donné naissance au modèle Fisher –, il se révèle plus problématique dans les milieux des sciences sociales. Il peut donner des réponses claires à la mauvaise question et produire des données impossibles à utiliser pour répondre à des questions d'évaluation cruciales, même lorsqu'il est possible de répondre clairement à la question Q-1.

Preuves empiriques sur le biais de randomisation

Les violations de l'hypothèse (HY-1) en général rendent peu fiables les preuves empiriques issues des expérimentations sociales randomisées. Quelle est, dans la pratique, l'importance de cette possibilité théorique ? Étonnamment, on ne sait que très peu de choses sur la réponse à cette question pour les expérimentations sociales menées en économie. En effet, à l'exception d'un programme, l'expérimentation sociale randomisée n'a été mise en œuvre que sur des « projets pilotes » ou des « projets de démonstration » destinés à évaluer de nouveaux programmes inédits. Le risque de perturbation pour cause de randomisation ne peut être ni confirmé ni infirmé à l'appui des données de ces expérimentations. Dans un programme évalué par randomisation, la participation était obligatoire pour la population cible (Doolittle et Traeger, 1990). Partant, la randomisation n'a pas eu d'incidence sur le vivier de candidats, ni sur l'évaluation de l'admissibilité des candidats par les administrateurs du programme.

Heureusement, nous avons quelques informations disponibles sur cette question, même si elles sont indirectes. En réponse à la grande variabilité des estimations de l'impact des programmes de main-d'œuvre dérivées d'estimateurs non expérimentaux de Lalonde (1986) et de Fraker et Maynard (1987), le département américain du Travail a financé une évaluation expérimentale à grande échelle de la loi sur les partenariats en matière de formation professionnelle (*Job Training Partnership Act*, JTPA), qui était alors le principal dispositif pour la formation

du gouvernement aux États-Unis. Une évaluation randomisée a donc été mise en œuvre sur différents sites. L'organisation en charge de la mise en œuvre de cette expérimentation – la MDRC – est un défenseur ardent et efficace de la randomisation comme méthode d'évaluation des programmes sociaux.

Un rapport de cette organisation (DOOLITTLE et TRAEGER, 1990) fournit quelques informations autorisant une analyse approximative des préférences révélées[7]. À la fin des années 1980 et au début des années 1990, la formation professionnelle aux États-Unis était organisée par le truchement de centres géographiquement décentralisés. Ces centres recevaient des primes d'incitation dès lors qu'ils parvenaient à trouver des emplois « à haute rémunération » à des chômeurs et à des personnes bénéficiant de l'aide sociale. La participation des centres à l'expérimentation n'était pas obligatoire. Une réserve a été constituée pour compenser les coûts administratifs de l'expérimentation à la charge des centres de formation professionnelle. Les fonds de cette réserve représentaient 5 à 10 % des coûts de fonctionnement totaux des centres.

En essayant d'enrôler des sites géographiquement dispersés, la MDRC a enregistré un taux de refus des centres de formation de plus de 90 %. Les raisons du refus de participation sont indiquées dans le tabl. 1 (Les raisons énoncées ne sont pas mutuellement exclusives). En tête de liste figurent des objections éthiques et des objections en termes de relations publiques. Des craintes majeures (items 2 et 3) ont été exprimées quant aux effets de la randomisation sur la qualité du vivier de candidats, qui entraverait la rentabilité des centres de formation. Avec la randomisation, les centres ont dû élargir le vivier de personnes jugées éligibles, et les effets de cet élargissement sur la qualité des candidats – à savoir le comportement exclu par l'hypothèse (HY-1) – ont suscité de vives inquiétudes. Afin d'inciter les centres à participer, la MDRC a dû réduire la probabilité de rejet aléatoire, passée de 1/2 à un niveau aussi bas que 1/6 pour certains centres. La taille réduite de l'échantillon de contrôle qui en résulte nuit à la puissance des tests statistiques conçus pour tester l'hypothèse nulle d'absence d'effet du programme. La compensation a été multipliée par sept afin que tous les centres puissent participer à l'expérience. Les analystes de la MDRC ont formulé les conclusions suivantes :

> « La mise en œuvre d'un plan de recherche complexe par assignation randomisée dans un programme en cours fournissant une variété de services modifie inévitablement son fonctionnement, d'une manière ou d'une autre […] La différence la plus probable découlant d'une étude de terrain par assignation randomisée des impacts du programme […] est un changement dans la composition de la clientèle servie. Les efforts accrus de recrutement, nécessaires pour constituer le groupe de contrôle, attirent des candidats supplémentaires qui ne sont pas identiques aux personnes précédemment servies. Un deuxième changement probable est que les catégories de traitement peuvent quelque peu restreindre la

7. HOTZ (1992) résume également leur discussion.

flexibilité du personnel du programme à modifier les recommandations du service » (DOOLITTLE et TRAEGER, 1990 : 121).

Ces auteurs ajoutent que « certains [centres de formation], en raison de graves problèmes de recrutement ou de services de première ligne, ne peuvent pas mettre en œuvre le type de modèle d'assignation randomisée nécessaire pour répondre aux différentes questions d'impact sans apporter de changements majeurs aux procédures » (*ibid.* : 123).

Tableau 1
Pourcentage d'agences locales JTPA citant des préoccupations spécifiques sur la participation à l'expérimentation.

	Préoccupation	Pourcentage de centres de formation citant la préoccupation
1.	Implications éthiques et de relations publiques :	
	a. de l'assignation aléatoire aux programmes sociaux	61,8
	b. du déni de services au groupe de contrôle	54,5
2.	Effet négatif potentiel de la création d'un groupe de contrôle sur la réalisation des objectifs de recrutement des clients	47,8
3.	Impact négatif potentiel sur les normes de performance	25,4
4.	Mise en œuvre de l'étude lorsque les prestataires de services procèdent au recrutement	21,1
5.	Objections des prestataires de services à l'égard de l'étude	17,5
6.	Charge administrative potentielle pour le personnel	16,2
7.	Manque de soutien possible de la part des élus	15,8
8.	Légalité de l'assignation aléatoire et griefs éventuels	14,5
9.	Procédures pour orienter le groupe de contrôle vers d'autres services	14,0
10.	Problèmes particuliers de recrutement pour les jeunes non scolarisés	10,5
	Taille de l'échantillon	228

Source : réponses de 228 agences JTPA locales contactées au sujet d'une éventuelle participation à l'étude nationale JTPA (DOOLITTLE et TRAEGER, 1990).

Crédits : Manpower Demonstration Research Corporation.

Notes : les préoccupations relevées par moins de 5 % des centres de formation ne sont pas mentionnées. La somme des pourcentages peut dépasser 100, car les centres de formation peuvent citer plus d'une préoccupation.

Durant l'expérimentation conduite à Corpus Christi, au Texas, les administrateurs du centre ont demandé et obtenu du gouvernement du Texas de pouvoir déroger à ses normes de performance au motif que l'expérimentation avait perturbé les opérations du centre. L'auto-sélection réduit probablement le risque de perturbation pour les sites participants. Une telle participation sélective à l'expérimentation remet en question la validité des estimations expérimentales en ce qui concerne le système JTPA dans son ensemble. Au moins, les données peuvent être utilisées pour établir une estimation plus précise de l'impact majeur de la perturbation.

La randomisation est également controversée dans les essais cliniques en médecine, qui sont parfois présentés comme un parangon pour les sciences sociales empiriques (voir par exemple ASHENFELTER et CARD, 1985). Le problème éthique soulevé par les centres de formation de main-d'œuvre, qui consiste à refuser l'accès à la formation à des personnes également qualifiées, trouve son pendant dans l'application des essais cliniques randomisés. Par exemple, PALCA (1989), dans la revue *Science*, note que des patients atteints du Sida, privés de médicaments potentiellement vitaux, ont pris des mesures pour neutraliser l'assignation randomisée. Les patients ont fait tester les pilules qui leur étaient administrées pour voir s'ils recevaient un placebo ou un traitement insatisfaisant. Dans un cas comme dans l'autre, ils étaient susceptibles d'abandonner l'expérience ou de rechercher un médicament plus efficace, ou les deux. Pour ce qui est de l'expérimentation de la MDRC, les stagiaires qualifiés de certains sites ont trouvé d'autres moyens d'obtenir exactement la même formation dispensée par les mêmes sous-traitants, avec d'autres mécanismes de soutien financier.

Dans le *Journal of the American Medical Association*, KRAMER et SHAPIRO (1984 : 2739) notent que les sujets d'essais pharmaceutiques sont moins susceptibles de participer à des essais randomisés qu'à des études non expérimentales. Ils évoquent en particulier une étude sur des médicaments administrés à des enfants atteints d'une maladie. L'étude comportait deux volets. La phase non expérimentale de l'étude a connu un taux de refus de 4 %, tandis que 34 % d'un sous-échantillon des mêmes parents ont refusé de participer à un sous-essai randomisé, alors que les traitements proposés étaient tout aussi dénués de risques.

Ces auteurs invoquent des preuves empiriques suggérant que la non-réponse à la randomisation est sélective. Dans une étude sur le traitement d'adultes atteints de cirrhose, aucun effet du traitement n'a été constaté pour les participants à un essai randomisé. Mais les taux de mortalité des personnes randomisées et ne recevant pas le traitement étaient sensiblement inférieurs à ceux des personnes qui ont refusé de participer à l'expérience, bien que les deux groupes aient reçu le même traitement alternatif.

Cette preuve nuance le plaidoyer en faveur de l'expérimentation sociale randomisée. Lorsque celle-ci est possible, elle peut modifier le programme étudié. Pour de nombreux programmes sociaux, ce n'est pas un outil d'évaluation réaliste.

À quel stade la randomisation doit-elle être mise en œuvre ?

Jusqu'ici, j'ai délibérément fait abstraction de la caractéristique multi-étapes de la plupart des programmes sociaux. Dans cette section, j'examine brièvement la question du choix du stade d'un programme en plusieurs étapes auquel la randomisation devrait être mise en œuvre.

En principe, une randomisation pourrait être effectuée pour évaluer les résultats à chaque étape, mais la rareté des randomisations multiples indique probablement qu'elles exacerberaient le problème du biais de randomisation discuté dans les deux sections précédentes. En supposant qu'il n'y ait pas de biais de randomisation, si une seule randomisation doit être effectuée, à quel stade doit-elle l'être ? Une réponse évidente est le stade où elle est la moins perturbatrice, même si ce stade n'est pas si facile à déterminer en l'absence d'informations substantielles sur le processus étudié. Si la randomisation est effectuée à un stade donné, des estimateurs « économétriques » ou « statistiques » non expérimentaux sont nécessaires pour évaluer les résultats attribuables à la participation à tous les autres stades. Cela explique les analyses parfois très compliquées (HAM et LALONDE, 1990) ou controversées (CAIN et WISSOKER, 1990 ; HANNAN et BRANDON, 1990) des données expérimentales randomisées qui sont apparues dans la littérature récente.

En outre, pour certaines des questions posées au début du papier, il n'est pas évident que la randomisation soit la méthode à préconiser pour garantir des réponses convaincantes. Nombre des questions énumérées ici concernent la réponse des stagiaires et des centres de formation aux variations des contraintes. Si une variation accrue des variables explicatives (dans un sens précisé par CONLISK, 1973) facilite l'estimation des fonctions de réponse, il n'y a aucune raison pour que les allocations randomisées soient souhaitables ou optimales à cette fin.

Ainsi, si nous cherchons à améliorer nos connaissances sur la façon dont le revenu familial détermine la participation au programme, il n'est pas évident que les allocations de suppléments de revenu familial assignées de façon aléatoire soient un substitut efficient ou optimal aux stratégies non expérimentales de plan d'échantillonnage optimal qui suréchantillonnent le revenu familial aux extrêmes de la population éligible[8].

Si nous cherchons à améliorer nos connaissances sur la manière dont les conditions du marché du travail affectent l'enrôlement et le mantien dans le programme, ou encore les décisions d'acceptation ou de placement par les centres de formation, il serait souhaitable d'avoir une variation entre les sites de formation et ces conditions. Il n'est pas évident que la randomisation soit le meilleur moyen de garantir cette variation.

La randomisation de l'éligibilité au programme a été proposée comme alternative à la randomisation lors de l'inscription. Cette modalité est parfois considérée comme un type de randomisation plus acceptable, car il évite les coûts de candidature et de sélection encourus lorsque la randomiation exclut les personnes acceptées du programme. Étant donné que la randomisation est effectuée en dehors du centre de formation, le centre n'a pas à assumer le coût politique induit lorsqu'il refuse à des personnes éligibles le droit de participer

8. Cette remarque suppose un modèle linéaire. Pour des plans optimaux dans des modèles non linéaires, voir par exemple SILVEY (1980).

au programme. C'est pourquoi la randomisation est peut-être moins disruptive lorsqu'elle est effectuée à un autre stade.

Si l'éligibilité est assignée de manière aléatoire dans la population, elle se heurte toujours au problème de l'auto-sélection. En supposant que l'éligibilité ne perturbe pas les paramètres fondamentaux du programme, le simple paramètre de différence moyenne comparant les personnes éligibles aux personnes inéligibles identifie $E\left(Y_1 - Y_0 \middle| D = 1\right)P$, où P est la probabilité de participation au programme par sélection volontaire en l'absence d'expérimentation. En divisant par P, on peut identifier le traitement sur les personnes traitées.

La tension entre les expérimentations sociales comme substitut aux modèles comportementaux et les expérimentations sociales comme source d'information supplémentaire

Il existe une tension intellectuelle entre le point de vue du plan expérimental optimal et le point de vue de la différence moyenne simple à l'égard des expérimentations sociales. Le point de vue du plan expérimental optimal, plus ancien, privilégie des modèles explicites et utilise les expérimentations pour obtenir des paramètres de modèles comportementaux ou « structurels ». Le point de vue de la randomisation simple cherche à contourner les modèles et donne – sous certaines conditions – une réponse nette à une question (Q-1) : le programme fonctionne-t-il pour les participants ? Les deux points de vue peuvent être conciliés si l'on fait preuve d'agnosticisme quant aux informations préalables dont disposent les analystes pour concevoir les expérimentations (SAVAGE, 1962). Cependant, les avantages de la randomisation sont moins évidents lorsque l'objectif est de récupérer des fonctions de participation et de poursuite des stagiaires que s'il s'agit de récupérer la distribution des mesures de résultats du programme.

Le conflit potentiel entre les objectifs de l'expérimentation comme moyen d'obtenir de meilleures estimations d'un modèle comportemental et l'expérimentation comme méthode de production d'estimateurs simples des impacts moyens d'un programme apparaît avec force lorsque nous envisageons d'utiliser des données provenant d'expérimentations randomisées pour estimer un modèle comportemental. Pour se concentrer sur les points principaux, prenons l'exemple d'un programme en deux étapes. $D_1 = 1$ si une personne passe la première étape ; = 0 sinon. $D_2 = 1$ si une personne passe la deuxième étape ; = 0 sinon. Supposons que le résultat Y puisse être écrit sous la forme suivante :

$$Y = \theta_0 + \theta_1 D_1 + \theta_2 D_1 D_2 + U. \qquad (14)$$

Le problème statistique est que D_1 et D_2 sont stochastiquement dépendants de U. La randomisation à la première étape rend D_1 indépendant de U. Elle ne garantit pas que $D_1 D_2$ soit stochastiquement indépendant de U.

L'estimateur simple de différence moyenne, qui compare les résultats des personnes ayant achevé la première étape avec les résultats des personnes randomisées, estime, dans de grands échantillons :

$$E\left(Y|D_1=1\right)-E\left(Y|D_1=0\right)=\theta_1+\theta_2 E\left(D_2\middle|D_1=1\right).$$

Afin d'estimer θ_2 ou θ_1 pour mesure les effets marginaux de l'achèvement du programme à chaque étape, il faut trouver une variable instrumentale pour $D_1 D_2$.

La randomisation sur une seule coordonnée élimine uniquement la nécessité d'avoir un seul instrument pour accomplir cette tâche. La question de savoir à quelle étape la randomisation doit être mise en œuvre demeure une question ouverte. Le compromis entre la randomisation en tant que variable instrumentale et un meilleur plan d'échantillonnage non expérimental reste à étudier. Le plan optimal d'une expérimentation pour estimer les paramètres de (14) en général n'impliquerait pas une randomisation simple à une étape particulière. Les données générées comme sous-produit d'une randomisation simple ne sont idéales que pour l'estimation de modèles tels que (14), dans le sens limité où il faut une variable instrumentale de moins pour estimer de manière cohérente θ_1 ou θ_2, bien que ces données constituent un réel avantage.

Résumé du papier de 1992

Ce chapitre examine de manière critique les arguments avancés lors du Premier Réveil en faveur de l'expérimentation sociale randomisée comme méthode d'évaluation de programmes sociaux. La méthode donne des réponses convaincantes à certaines questions politiques en reposant sur des hypothèses fortes concernant le comportement des agents et les questions d'intérêt pour les évaluateurs de programmes.

La méthode est idéale pour évaluer des programmes sociaux si l'on se focalise sur l'estimation de l'effet *moyen* du traitement sur les résultats des personnes traitées et si l'une des hypothèses suivantes se vérifie :

> **(HY-1),** la randomisation n'a aucun effet sur les décisions de participation ;

ou

> **(HY-2),** s'il y a un effet de la randomisation sur les décisions de participation, soit (a) l'effet du traitement est le même pour tous, soit (b) si les agents diffèrent dans leurs réponses aux traitements, leurs réponses

idiosyncrasiques au traitement n'influencent pas leurs décisions de participation.

Si l'on se focalise sur d'autres caractéristiques des programmes sociaux, comme les déterminants des décisions de participation, de rejet ou de poursuite, les données randomisées ne présentent aucun avantage comparatif par rapport à des données stratifiées non randomisées. Même si (HY-1) se vérifie, il est impossible d'employer des données expérimentales pour étudier la distribution des résultats du programme ou leur médiane sans formuler des hypothèses « statistiques » ou « économétriques » supplémentaires. Dans un programme en plusieurs étapes, les données expérimentales randomisées produisent un estimateur « propre » (différence moyenne) de l'impact du programme uniquement pour les résultats définis de manière conditionnelle à l'étape ou aux étapes où la randomisation est mise en œuvre. Il faut malgré tout employer des méthodes statistiques, et les hypothèses qui les accompagnent, pour évaluer les résultats à d'autres étapes et les résultats marginaux à chaque étape.

Avec des hypothèses garantissant des réponses valables, la méthode expérimentale randomisée contourne la nécessité de spécifier des modèles comportementaux élaborés. Cependant, les preuves expérimentales acquièrent alors une certaine rigidité pour prédire les résultats dans des environnements différents de ceux utilisés pour mener l'expérience. L'interpolation et l'extrapolation remplacent les prévisions basées sur des modèles. Cela étant, de telles procédures d'ajustement de la courbe peuvent produire des prévisions plus convaincantes que celles générées à partir d'un modèle comportemental controversé.

L'hypothèse (HY-1) n'est pas controversée dans le contexte de l'expérimentation agricole randomisée. C'est dans ce contexte que le modèle d'expérimentations de FISHER (1935) a été développé. Ce modèle est le fondement intellectuel des expérimentations sociales récentes, bien que la littérature économique l'attribue souvent à tort aux statisticiens des années 1970. L'hypothèse (HY-1) est toutefois plus controversée dans le contexte des essais cliniques en médecine. Des agents humains peuvent répondre à la randomisation et ces réponses menacent potentiellement la fiabilité des preuves expérimentales. Les preuves sur le biais de randomisation présentées précédemment remettent en question la validité de (HY-1).

Si cette hypothèse n'est pas valable, si les participants au programme réagissent différemment aux traitements communs et que ces différences déterminent au moins en partie les décisions de participation au programme – de sorte que (HY-2) est fausse –, les méthodes expérimentales n'estiment pas même l'effet moyen du traitement sur les personnes traitées. Dans ce cas, les méthodes expérimentales randomisées répondent à la mauvaise question, à moins que la randomisation ne soit une caractéristique permanente du programme social évalué. Les données issues d'expérimentations randomisées ne peuvent pas être employées pour estimer des équations de participation, d'inscription et de poursuite pour les programmes en cours.

Postscript, 2019

Je m'en tiens à ma discussion des questions conceptuelles soulevées dans ce chapitre et dans mon article avec Smith (Heckman et Smith, 1995)[9]. Les points soulevés sont tous valables aujourd'hui et ont été largement ignorés lors du récent Second Réveil observé en économie du développement. De nombreux documents rédigés à la suite de ces papiers établissent ou réitèrent les points soulevés ici. Outre qu'ils ne tirent pas les leçons du passé, les *randomistas* se montrent peu généreux envers les véritables pionniers des expérimentations de terrain.

Dans des travaux ultérieurs, Heckman et Smith (1998) développent le point selon lequel l'auto-sélection dans un programme génère des informations sur les perceptions *ex ante* des agents quant aux avantages du programme[10]. Ces évaluations subjectives sont sans doute plus importantes que les évaluations « objectives » (δ) privilégiées par les statisticiens qui considèrent le non-respect du protocole comme un problème plutôt que comme une source d'information. Ces informations disparaîtraient si les personnes étaient forcées de s'inscrire dans un groupe de traitement ou dans un groupe de contrôle. Ce point est un exemple de plus des avantages qu'il y a à se servir de l'économie pour concevoir et évaluer des programmes sociaux.

Dans des travaux ultérieurs, Heckman *et al.* (2000) voient dans le *biais de substitution* une menace majeure à l'interprétation directe des expérimentations. Si les agents ont accès à des programmes alternatifs, les personnes éligibles à un programme et les personnes non éligibles peuvent choisir de participer à un programme alternatif. La différence moyenne « transparente » entre groupe de traitement et de contrôle ne compare pas l'effet du traitement avec l'absence de traitement, mais plutôt l'effet du traitement par rapport à la meilleure alternative, qui peut en fait être meilleure que le programme évalué. Notre papier (Heckman *et al.*, 2000) documente l'omniprésence du problème. Kline et Walters (2016) ont récemment livré une démonstration du problème du biais de substitution. L'estimateur « transparent » de la différence moyenne d'une récente évaluation expérimentale du programme Head Start suggérait que le programme n'avait aucun impact sur les enfants défavorisés. Une analyse plus approfondie tenant compte du biais de substitution à l'aide de méthodes microéconométriques montre un effet important. Leur article fait écho aux conclusions de Cogan (Congress of the United States, SCOFSOPA, 1978) concernant l'INR.

Banerjee et Duflo (2009) répondent aux points soulevés dans mon papier de 1992, tout comme Athey et Imbens (2017). Ils affirment que ces critiques n'ont plus lieu d'être en raison de l'amélioration des plans d'enquête et de la méthodologie de mise en œuvre. Toutefois, ils n'abordent pas de nombreux points interprétatifs ou conceptuels de base soulevés dans mon papier de 1992

9. J'ai depuis amplifié ces points dans Heckman *et al.* (1997b ; 1999) et Heckman et Vytlacil (2007).

10. Ainsi, comme le notent Heckman et Smith (1998), la douleur et la souffrance d'un essai médical peuvent l'emporter sur ses avantages en termes de survie.

ou dans l'article de 1995, ni l'incapacité des comparaisons expérimentales des différences moyennes à répondre à l'éventail des effets de traitement pertinents au plan politique qui sont discutés dans mes papiers et dans des études ultérieures (HECKMAN, 2008).

La littérature postérieure à mon papier de 1992 a fourni des preuves considérables de l'insuffisance des preuves expérimentales dans de nombreux domaines. SANSON-FISHER *et al.* (2007) montrent que les expérimentations ont une portée fondamentalement trop limitée pour prendre en compte les évaluations d'impact, comme l'autonomisation des femmes. CONCATO et HORWITZ (2018) étudient le consensus en médecine[11]. Ce consensus ne voit plus dans les RCT un « étalon-or », une réputation qui, selon eux, était de mise dans les années 1990 en médecine. Ils présentent de nombreux articles sur les limites des expérimentations randomisées en médecine (HORWITZ, 1996 ; FEINSTEIN et HORWITZ, 1997 ; SHAHAR, 1997 ; SEHON et STANLEY, 2003 ; CONCATO et HORWITZ, 2004 ; CHAKRAVARTY et FRIES, 2006 ; WORRALL, 2007 ; RAWLINS, 2008 ; BORGERSON, 2009 ; CONCATO, 2012 ; 2013 ; HORWITZ et SINGER, 2017 ; FRIEDEN, 2017). CZIBOR *et al.* (2019) ont récemment publié un article de mise en garde pour les économistes expérimentaux, qui reprend les points de mon papier de 1992. Cet article met en lumière les graves problèmes de l'économie expérimentale et ce dont les fervents expérimentateurs doivent se méfier.

Les modèles causaux préconisés dans la récente littérature sur l'évaluation des programmes sont motivés par l'idée de l'expérimentation comme idéal. Ils ne précisent pas clairement les mécanismes théoriques déterminant les ensembles de résultats contrefactuels possibles, la manière dont les contrefactuels hypothétiques sont réalisés ou le mode de mise en œuvre des interventions hypothétiques, si ce n'est pour comparer des interventions « randomisées » avec des interventions « non randomisées ». Ils se concentrent sur les résultats, sans spécifier le modèle de sélection des résultats ni les préférences des agents par rapport aux résultats attendus (HECKMAN, 2008).

Quiconque ignore l'histoire intellectuelle est condamné à répéter les erreurs du passé. La seconde vague se retirera lorsque les économistes réapprendront les leçons du passé.

11. Le « parangon » cité par ASHENFELTER et CARD (1985).

Entretiens

Entretien avec Jean-Paul Moatti et Rémy Rioux

Rémy Rioux, vous dirigez une institution (l'Agence française de développement – l'AFD) dont la mission est de financer des projets, des programmes et des politiques de développement. En quoi consiste l'évaluation dans votre institution et quel rôle joue-t-elle dans votre activité et celle de vos partenaires ?

Pour bien comprendre le rôle de l'évaluation, il convient d'abord de la situer dans le cycle de travail de l'Agence française de développement à la lumière de ce que j'appelle, dans mon livre *Réconciliations*, l'approche des 4E pour *Écoute*, *Expertise*, *Engagement* et *Évaluation*. Le point de départ de l'identification par l'AFD d'un projet ou d'une politique publique est toujours l'expression d'un besoin par un partenaire d'un pays du Sud, comme un ministère, une collectivité locale ou un organisme de coopération internationale. Si ce besoin est en phase avec les priorités d'aide au développement qui ont été fixées à notre agence par le gouvernement français, nous débutons alors une phase d'écoute et de dialogue dans laquelle l'AFD s'appuie sur des expertises internes et externes pour analyser l'opportunité du projet, ainsi que sa faisabilité sur les plans technique, financier et institutionnel. Cette phase de conception est essentielle, car elle conditionne le champ de l'évaluation du projet en définissant, avec les parties prenantes, les objectifs de transformation poursuivis par le projet.

Le projet est ensuite passé au crible d'un « avis développement durable », établi par une structure indépendante des opérations, qui note le projet selon une grille d'analyse composite afin d'estimer les impacts potentiels sur six dimensions du développement durable. L'évaluation ne porte pas seulement sur les réalisations matérielles, mais aussi – et surtout – sur l'atteinte de l'objectif du projet. Nos actions de développement doivent en effet être évaluées par rapport à ce qu'elles apportent au tissu économique, environnemental et social local. Par exemple, si un projet vise à augmenter le taux de scolarisation des enfants et à améliorer leurs acquis éducatifs, la construction d'établissements fonctionnels et pérennes est une condition importante, mais non suffisante pour le succès

du projet. Le projet ne sera considéré comme réussi que si ces bâtiments sont fortement fréquentés et que l'on constate une amélioration significative des compétences des élèves en fin de cursus scolaire. Par conséquent, l'évaluation n'est possible que si la logique du changement a été clairement définie dès la conception du projet et si des critères de réussite, qui se traduisent par des indicateurs quantitatifs et qualitatifs renseignés de manière méthodique, ont été explicitement identifiés dès le départ.

Enfin, un dispositif de suivi doit être mis en place avant de démarrer la mise en œuvre du projet afin de contrôler non seulement le bon déroulement des activités financées, mais aussi l'atteinte des résultats intermédiaires. Si l'on reprend l'exemple de notre projet éducatif, cela signifie que l'on doit s'assurer, avant qu'il ne touche à sa fin, que les enseignants ont été formés et affectés dans les établissements construits dans le cadre du projet, et que les inégalités d'accès entre les filles et les garçons se réduisent. Si ce n'est pas le cas, il faudra prendre des mesures correctives avant l'achèvement du projet pour atteindre les objectifs fixés.

L'évaluation joue donc un rôle éminemment opérationnel. C'est un élément clé du métier de développeur pour comprendre les paramètres qui ont permis, accéléré ou freiné le déroulement du projet et la réalisation de ses objectifs, en tirer des enseignements et être plus efficace. Mais l'évaluation joue également un rôle de plus en plus stratégique. À l'heure où les besoins d'investissements pour réduire les inégalités et préserver les équilibres de notre planète vont croissant, la question principale qu'un bailleur de fonds comme l'AFD se pose est la suivante : comment l'évaluation peut-elle nous servir à coconstruire des projets de développement avec nos partenaires, et à soutenir des politiques publiques plus efficaces et efficientes au bénéfice des populations ? L'une des leçons que je tire de notre expérience à l'AFD est qu'évaluer notre action, c'est s'inscrire dans la démarche de redevabilité et d'amélioration continue. Voilà pourquoi l'évaluation joue un rôle central et croissant dans notre institution.

Il faut reconnaître qu'en raison de son histoire, l'évaluation de l'aide française a longtemps été reléguée au second plan par rapport aux autres pays et aux bailleurs multilatéraux. Les études comparatives de Laporte (2015) montrent que les pratiques d'évaluation ont émergé parallèlement en France et dans les pays anglo-saxons dans les années 1960, avec une approche spécifique en France, marquée par le rôle de l'Institut national de la statistique et des études économiques (Insee) et des experts en agronomie et développement rural de la coopération française. La France a toutefois pris du retard dans les années 1980 et 1990, à une époque où les bailleurs anglo-saxons systématisaient l'évaluation à mesure qu'augmentaient leurs budgets consacrés à l'aide internationale.

La France rattrape maintenant ce retard sous l'effet conjoint d'investissements croissants consacrés au développement durable et des exigences de résultats qui y sont associées. Les moyens dont dispose l'AFD augmentent progressivement pour atteindre l'objectif fixé par le Président français de consacrer 0,55 % du PIB à l'aide publique au développement d'ici 2022. Ces moyens accrus

s'accompagnent naturellement d'un impératif : celui de rendre toujours mieux compte, au gouvernement, aux parlementaires et aux citoyens, de l'efficacité de notre action. L'évaluation joue un rôle déterminant pour répondre à cette exigence de redevabilité.

*Parmi ses évaluations, l'AFD mène des études d'impact, qui ont une ambition scientifique plus poussée. Quelle est l'utilité de ces recherches évaluatives, et que pensez-vous de l'évaluation par assignation aléatoire (*Randomized Controlled Trial – *RCT), qui est présentée aujourd'hui comme la méthode la plus rigoureuse en la matière ?*

Dans ma réponse précédente, je me référais aux évaluations en général. Les études d'impact constituent un type particulier d'évaluation, qui est conçu pour identifier, mesurer et comprendre de manière scientifiquement rigoureuse les effets strictement imputables à une intervention. Elles s'appuient pour cela sur un contrefactuel, c'est-à-dire qu'elles mettent en regard l'évolution d'une population bénéficiaire à celle d'une population non bénéficiaire, en s'assurant que ces deux populations sont effectivement comparables et que l'intervention évaluée est le seul critère qui les différencie.

Les évaluations d'impact sont essentiellement utilisées comme un outil ponctuel pour démontrer (*to prove*), plutôt que comme un outil d'amélioration (*to improve*). Elles permettent d'établir qu'en général certains types d'interventions fonctionnent dans des contextes donnés, mais elles s'avèrent trop lourdes, longues et coûteuses pour être systématisées comme un outil de redevabilité et d'apprentissage. Pour ceux-ci, on dispose de méthodes plus légères et harmonisées, qui ont été régulièrement améliorées au cours des trois dernières décennies. Ces démarches peuvent s'appuyer sur de l'expertise locale existante et être mobilisées de manière plus agile pour alimenter le dialogue avec nos partenaires.

L'AFD réalise des études d'impact depuis le début des années 2000 pour participer à l'enrichissement d'un corpus de connaissances générales sur le développement. L'analyse comparative détaillée que nous avons élaborée sur ce sujet (Pamiès-Sumner, 2015) montrait qu'il s'agissait d'une particularité de notre institution. À l'exception du Department for International Development (DFID), la plupart des bailleurs bilatéraux restent en retrait en matière d'évaluations d'impact par rapport aux grands bailleurs multilatéraux qui en ont produit des centaines.

Depuis quinze ans, l'AFD a également contribué activement à la réflexion méthodologique de la communauté du développement sur les évaluations d'impact. Parmi les méthodes qui peuvent être utilisées pour renforcer la comparabilité entre population bénéficiaire et contrefactuel pour déduire l'impact d'une intervention, la méthode des RCT a connu un fort engouement sur lequel le présent ouvrage se propose de prendre du recul. L'AFD a joué un rôle clé dans l'amorce de cette tendance en soutenant, dès 2005, deux vastes RCT sur des secteurs d'intervention qui étaient alors au cœur de l'actualité de l'aide : le microcrédit

et l'assurance santé. Ces études nous ont permis de faire progresser l'état des connaissances sur ces deux secteurs d'intervention, mais aussi sur les méthodes expérimentales, dont nous avions alors tiré un bilan mitigé (BERNARD *et al.*, 2012). Avec ces travaux, nous avons participé à l'émergence d'un consensus sur la nécessité d'évoluer vers des approches pluridisciplinaires combinant méthodes quantitatives et qualitatives.

Aujourd'hui, il est plus indispensable que jamais de mener des actions efficaces. Il ne reste que neuf ans pour atteindre les Objectifs de développement durable (ODD). Les interventions publiques doivent s'appuyer sur des connaissances empiriques si l'on veut répondre aux exigences légitimes des citoyens, dans un monde déchiré par des fractures sociales, politiques, économiques et, bien sûr, environnementales. En tant que plateforme pour le développement durable, l'AFD se doit d'être pragmatique et efficiente, et nous sommes déterminés à réaliser davantage d'évaluations d'impact. Nous avons les moyens, objectifs et scientifiques, de savoir ce qui fonctionne et ce qui ne fonctionne pas.

Il reste une marge de progression. Comme je le souligne dans *Réconciliations*, les approches actuelles d'évaluation de l'impact de l'aide au développement sont loin d'être parfaites. Les résultats sont généralement mesurés par un instantané, utile, mais insuffisant, d'indicateurs quantitatifs tels que le nombre de personnes raccordées au réseau électrique, le taux de jeunes filles scolarisées ou les tonnes de CO_2 économisées. Les approches quantitatives sont souvent décevantes, car elles ne permettent pas d'identifier dans quelle mesure un projet de développement contribue aux dimensions qualitatives de l'agenda 2030. À cet égard, il est essentiel de définir un cadre commun pour déterminer les investissements qui sont en phase avec les trajectoires durables à long terme et ceux qui ne le sont pas, afin d'améliorer les approches qualitatives de l'évaluation pour atteindre les objectifs de l'accord de Paris et des ODD.

Comme les méthodologies se sont diversifiées, l'important est maintenant d'utiliser les approches les plus pertinentes pour répondre aux questions sur les sujets étudiés, sans préjugés théoriques. C'est la raison pour laquelle les équipes de l'AFD ont le souci constant de combiner les études d'impact et les évaluations opérationnelles plus légères en développant un ensemble diversifié d'outils de mesure permettant de mieux appréhender l'efficacité de l'aide. Nous prévoyons en particulier de mettre davantage la science au service de l'évaluation, en collaborant avec des partenaires comme l'Institut de recherche pour le développement (IRD) et le monde scientifique en général, notamment par la promotion des recherches menées par nos partenaires dans les pays du Sud.

Jean-Paul Moatti, vous étiez, jusqu'à une date récente, président-directeur général de l'IRD. Mais d'abord, en tant que spécialiste en économie de la santé, vous connaissez bien les essais cliniques. Que pensez-vous des RCT à l'aune de votre propre expérience ?

J'ai consacré la plupart de mes quarante années de carrière universitaire à collaborer étroitement avec des épidémiologistes et des biostatisticiens qui ont

longtemps considéré les évaluations randomisées comme un outil essentiel pour rationaliser la pratique médicale sur la base de preuves scientifiques, mais qui ont fini par exprimer certaines préoccupations sur la confiance excessive accordée aux protocoles randomisés.

On a longtemps pensé que les RCT constituaient la source idéale (le « *Gold Standard* », « l'étalon-or ») de données sur les effets des traitements médicaux. D'autres méthodes, comme les études de cohorte et les études cas-témoins, sont bien entendu utilisées lorsque la randomisation n'est pas possible pour des raisons éthiques ou pratiques, comme c'est souvent le cas des études sur les facteurs de risque environnementaux. Ces dernières années s'est néanmoins développé un intérêt croissant pour d'autres méthodes capables de fournir des preuves empiriques pour les stratégies les plus efficaces dans les domaines de la médecine et de la santé publique.

Des RCT bien conçues peuvent bien sûr prétendre à une forte validité interne dans la mesure où elles répartissent de manière équilibrée les facteurs connus et inconnus entre les groupes de contrôle et de traitement, limitant ainsi le risque d'introduire des facteurs de confusion dans l'identification d'un mécanisme causal. Pourtant, les experts en santé publique reconnaissent depuis longtemps les limites des RCT sur le plan de leur validité externe et de leur pertinence pour la prise de décision. En effet, il peut y avoir un certain nombre de raisons qui font qu'une RCT manque de validité externe, et la généralisation des résultats hors de la population étudiée peut s'avérer erronée. Les RCT ne portent généralement pas sur des périodes d'étude suffisamment longues ou sur des populations assez importantes pour pouvoir évaluer, dans la durée, les effets d'un traitement, comme dans le cas de l'impact des vaccins sur l'immunité à long terme d'une population, ou pour mettre en évidence des effets indésirables rares, mais graves du médicament, qui apparaissent seulement lors des phases de surveillance après la mise sur le marché au travers des dispositifs de pharmacovigilance. Les contraintes croissantes de coût et de temps pesant sur les RCT peuvent inciter à se fier à des marqueurs de substitution potentiellement mal corrélés au résultat clinique (mortalité ou morbidité grave) que l'on cherche en réalité à évaluer. Pour limiter la dimension de l'échantillon et garantir une puissance statistique suffisante, les RCT se concentrent souvent sur les groupes à haut risque, ce qui réduit fréquemment leur pertinence pour des populations cibles plus larges. En outre, il faut des années pour planifier, mettre en œuvre et analyser la plupart des RCT, ce qui limite leur capacité à suivre le rythme des innovations biomédicales et oblige à prendre des décisions concernant les nouveaux médicaments et les dispositifs médicaux alors que leur évaluation clinique est encore en cours. Les contraintes de temps associées aux RCT limitent également leur utilisation efficace pour les décisions de santé publique en cas de flambées épidémiques ou de crises sanitaires. En outre, des RCT différentes traitant d'un même sujet ont souvent produit des résultats contradictoires, notamment en ce qui concerne des questions clés sur l'efficacité des pratiques médicales. Cela a conduit au développement de méthodes dites de *méta-analyse*, qui combinent les résultats de différents essais afin de surmonter ces contradictions et de parvenir à des

conclusions fondées. Mais ces méthodes soulèvent elles-mêmes des problèmes statistiques complexes ainsi que d'identification des essais dont les données peuvent être agrégées sans créer une hétérogénéité excessive.

De manière générale, les experts en santé publique tendent désormais à considérer que les systèmes actuels de hiérarchisation des preuves empiriques mises en œuvre dans les mécanismes d'évaluation des pratiques médicales sont biaisés en faveur des RCT et qu'ils sont susceptibles d'écarter des données valides qui ne sont pas issues d'expérimentations randomisées.

Plus récemment, un nombre croissant de biostatisticiens ont reconnu que la randomisation ne constitue pas en soi une garantie absolue de validité interne. Cook (2018) recense par exemple 26 hypothèses susceptibles de biaiser les résultats des RCT malgré la randomisation, 22 d'entre elles étant liées à la validité interne : une différence préexistante à la sélection des groupes de traitement et de contrôle qui pourrait être confondue avec un effet de traitement ; la possibilité que l'attribution ait pu varier d'un groupe à un autre dans l'essai, rendant les résultats très sensibles au traitement des données manquantes (notamment liés aux abandons et perdus de vue en cours d'essai) ; un biais dans le choix du groupe de contrôle (par exemple lorsque l'innovation est comparée à un standard actuel de traitement qui n'est pas optimal) ; des changements de comportement induits par la participation à l'essai (par exemple, dans certains essais en double aveugle avec un groupe de contrôle prenant un placebo, les patients infectés par le VIH ont commencé à partager leurs pilules afin de garantir que tous les participants reçoivent « au moins » quelques médicaments efficaces), etc. La plupart de ces préoccupations tournent autour d'une question clé : si la randomisation, qu'elle soit volontaire ou involontaire, qu'elle intervienne *ex ante* ou *ex post* (dans le mode d'analyse de l'essai), ignore les informations préalables issues de la théorie et la prise en compte des covariables, alors elle est susceptible d'introduire des biais, et peut même s'avérer contraire à l'éthique, car elle expose inutilement les participants à un danger éventuel dans le cadre d'une expérience risquée.

Des modifications de la structure de base de l'évaluation randomisée ont été élaborées pour minimiser ce risque grâce à des mesures comme la stratification, l'allocation adaptative et la pré-randomisation, visant à prévenir la distorsion entre les groupes, qui peut venir, en dépit du tirage au sort, de facteurs pronostiques connus ou identifiés. Les modèles recourant à la méthode du cas unique (*Single Case Design* ou SCD) sont utilisés lorsqu'une variable d'intérêt dépendante peut être mesurée de manière répétée dans le temps entre deux points (au départ et pendant ou après l'intervention). Plutôt que de recourir à la randomisation d'un grand nombre de participants, les chercheurs utilisent un ordonnancement minutieux et planifié des conditions expérimentales afin d'améliorer la validité interne en excluant toute autre explication des effets du traitement. Toutes ces tentatives d'amélioration de la validité interne constituent *de facto* un aveu que la randomisation ne présente, en pratique, aucune supériorité statistique intrinsèque pour l'inférence causale.

En mettant maintenant votre autre casquette, celle d'économétricien, vous connaissez sans doute bien les problématiques méthodologiques liées à l'inférence causale et les autres modèles de génération de groupes contrôle permettant d'identifier les facteurs causaux du phénomène d'intérêt.

Avec ces évolutions et la tendance actuelle vers l'« expérimentation pragmatique », les experts en santé publique ont redécouvert un « vieil » argument économétrique bayésien remontant à Fisher (1926) et Savage (1962), qui remet en question la croyance selon laquelle les effets de traitement moyens estimés à partir de RCT sont probablement plus proches de la vérité que ceux estimés par d'autres moyens. Le résultat visé par toute RCT réside dans l'écart de moyennes entre le groupe d'intervention et le groupe de contrôle, qui combine l'effet de traitement moyen estimé parmi les personnes traitées au terme d'erreur, qui lui-même traduit le déséquilibre généré aléatoirement parmi les effets nets issus d'autres causes. Les RCT offrent une base pour calculer l'ampleur de l'erreur, mais, comme indiqué précédemment, cela reste soumis à la condition qu'aucune corrélation avec des covariables ne soit intervenue avant ou après la randomisation. Que ce soit dans les RCT ou dans d'autres méthodes, la pertinence statistique sera menacée si la distribution des effets de traitement individuels dans la population étudiée est asymétrique. Dans sa synthèse des nombreuses publications économétriques sur le sujet, le prix Nobel d'économie Deaton soutient à juste titre que tout statut spécial accordé aux RCT est injustifié et conclut que « la méthode la plus à même de produire une bonne inférence causale dépend de ce que nous essayons de découvrir et de ce que nous savons déjà » (Deaton et Cartwright, 2018 : 2).

Les économètres connaissent d'autres méthodes conçues pour obtenir une inférence causale, comme l'appariement par scores de propension, les variables instrumentales, la modélisation économétrique et les réseaux bayésiens. Toutes les méthodes doivent bien entendu recourir à des groupes de contrôle pour garantir une comparaison appropriée avec le groupe d'intervention, mais le choix du modèle d'étude doit rester pragmatique et dépendre des problématiques en question.

Un autre prix Nobel d'économie, Heckman, met en lumière une autre source de scepticisme à l'égard des évaluations randomisées parmi les économètres, à savoir que les informations sur les effets de traitement moyens peuvent s'avérer peu utiles pour éclairer les politiques publiques, car elles ne prennent pas en compte les variations entre les bénéficiaires des interventions (Heckman et Smith, 1995). Les impacts moyens peuvent être le facteur d'intérêt principal dans une évaluation comparant deux médicaments ou deux interventions très simples, mais lorsqu'il s'agit de politiques à composantes multiples, la possibilité de tirer des enseignements utiles d'une expérience spécifique nécessite plutôt d'identifier les raisons pour lesquelles certaines stratégies fonctionnent mieux que d'autres. Dans de tels cas, même une RCT réussie ne peut garantir que la relation causale établie se maintiendra dans d'autres contextes ou de manière générale. Il paraît clairement fallacieux de prétendre que la microfinance doit

être au cœur des efforts pour éliminer la pauvreté ou que les transferts monétaires conditionnels doivent être la priorité des politiques en matière de santé et d'éducation en s'appuyant sur un nombre limité d'évaluations randomisées réalisées dans ces domaines. Cela peut détourner l'attention des grandes politiques publiques destinées à réduire les inégalités ou à assurer la santé et l'éducation pour tous. Il est intéressant de constater que certains partisans des évaluations randomisées, tels que BANERJEE *et al.* (2015b : 52), arrivent à une conclusion similaire lorsqu'ils reconnaissent, par exemple, que « les adeptes du microcrédit ont peut-être […] surestimé le potentiel des entreprises pour les pauvres, à la fois comme source de revenus et comme moyen d'autonomisation des femmes qui en sont les propriétaires ».

Rémy Rioux, avez-vous observé des retombées opérationnelles des évaluations au niveau des projets que vous soutenez sur le terrain, et quelles évolutions avez-vous remarquées ou souhaiteriez-vous voir ?

En général, nous constatons que la culture de l'évaluation favorise une culture de l'innovation. Les évaluations d'impact ont généré des retombées intéressantes dans la mesure où elles ont consolidé une culture de l'évaluation au sein de l'AFD et parmi nos partenaires des pays en développement.

En Mauritanie, par exemple, l'évaluation de l'impact d'un dispositif de protection sociale couvrant 40 % des femmes a montré que celui-ci augmentait significativement le recours aux soins et réduisait les inégalités, mais qu'il ne touchait pas les plus démunies et qu'il n'avait pas d'impact significatif sur la santé des mères et des enfants en raison de la dégradation de la qualité des soins dans les établissements (PHILIBERT *et al.*, 2017). En conséquence, nous avons complètement repensé la phase suivante du projet : action intégrale sur les différentes composantes de la qualité (ressources humaines, sang, médicaments et supervision) et opérationnalisation d'un mécanisme de gratuité pour les plus pauvres.

Les évaluations scientifiques d'impact ont également favorisé des innovations méthodologiques qui améliorent le suivi des projets. Par exemple, nous proposons désormais d'aider les responsables de projets et les partenaires à utiliser des données existantes dès l'instruction d'un projet pour mieux estimer les conditions de vie et l'accès aux services, analyser les dépenses des ménages, etc., ou encore à utiliser des images satellitaires pour suivre la productivité agricole, la déforestation, le développement urbain, etc. Le suivi continu des projets est également facilité par des outils numériques comme GeoPoppy, une application développée par l'Institut national de la recherche agronomique (Inra), que nous avons utilisée pour suivre l'agriculture en Côte d'Ivoire et que nous allons même faire évoluer en outil de renforcement des capacités au Bénin. Dans le même esprit, l'AFD collabore en Haïti et au Niger avec le Centre de recherches interdisciplinaires (CRI), fondé par François Taddei et Ariel Lindner, pour expérimenter et diffuser de nouvelles façons de mener des recherches et de mobiliser l'intelligence collective dans les domaines des sciences du vivant, de l'éducation et des technologies numériques.

Tous ces exemples montrent que, lorsque nous sommes à même de rendre compte des impacts de nos projets, nous apprenons à partager notre expérience, à mieux écouter nos bénéficiaires et, *in fine*, à innover pour eux et avec eux. En évaluant nos impacts, nous pouvons innover et montrer concrètement le retour sur investissement que génère la politique de développement, c'est-à-dire un investissement durable et solidaire.

*Jean-Paul Moatti, vous êtes aussi engagé dans la contribution aux politiques publiques. Outre votre parcours académique, vous avez été membre du groupe d'experts des Nations unies chargé du rapport mondial sur le développement durable (*Global Sustainable Development Report – GSDR*). À ce titre, vous avez participé à la rédaction du premier rapport quadriennal d'évaluation des ODD, qui ont été adoptés par tous les États membres des Nations unies en septembre 2015 et qui fixent l'Agenda international 2030 pour le développement. Nous assistons dans ce cadre à l'émergence du concept de « science de la durabilité ». Soutenez-vous ce concept et comment les expérimentations randomisées peuvent-elles, ou non, contribuer à une recherche efficace pour le développement durable ?*

Même si les 17 ODD restent le fruit de nombreux compromis entre les gouvernements et entre des intérêts contradictoires, ce programme de transformation ambitieux a très largement bénéficié de l'émergence de ce que l'on appelle aujourd'hui la « science de la durabilité ». L'Académie nationale des sciences des États-Unis, qui a commencé à promouvoir la science de la durabilité en 2000, la définit comme un « domaine de recherche émergent qui traite des interactions entre les systèmes naturels et sociaux, et de la manière dont ces interactions affectent le défi de la durabilité : satisfaire les besoins des générations actuelles et futures tout en réduisant sensiblement la pauvreté et en préservant les écosystèmes qui entretiennent la vie sur la planète » (KATES, 2011). Parce qu'elle est axée sur les problèmes, cette nouvelle approche scientifique est par nature interdisciplinaire et se prête à la coconstruction de programmes de recherche avec les communautés concernées. Elle s'attache à identifier les chaînes de causalité complexes qui génèrent les principales problématiques environnementales et sociales menaçant l'avenir de la Terre, et à proposer des solutions pour réduire le risque d'incohérence dans la mise en œuvre des ODD tout en maximisant les synergies positives qui existent entre eux : comment augmenter la productivité agricole pour garantir la sécurité alimentaire d'une population mondiale, censée augmenter de deux milliards de personnes d'ici 2050, tout en réduisant les intrants chimiques afin de limiter l'impact environnemental et le gaspillage des ressources ? Comment promouvoir une croissance durable pour éliminer l'extrême pauvreté sans accroître les inégalités à l'intérieur des pays ? Les enjeux sont multiples.

L'économie du développement devrait jouer un rôle de premier plan dans cette science de la durabilité, car elle mobilise des compétences et des connaissances essentielles pour la transposition en politiques publiques efficaces de faits et de preuves empiriques, issues d'une gamme étendue de disciplines allant des

sciences naturelles aux sciences sociales, afin de les adapter à des contextes sociaux, environnementaux et économiques hétérogènes. Les RCT ne sont bien souvent pas adaptées à ce champ interdisciplinaire essentiel, car l'extrapolation et la généralisation de leurs résultats nécessitent un ensemble d'informations complémentaires qui doivent provenir d'autres sources. Surestimer leur rôle, en ignorant les limites de l'approche randomisée et en exagérant leurs mérites auprès des décideurs, pourrait compromettre la contribution de la science économique à la transformation des modèles de développement actuels vers la durabilité. Cependant, ce serait également une erreur de sous-estimer le fait que les études randomisées peuvent s'avérer extrêmement utiles, le cas échéant, pour déterminer quelles sont les meilleures pratiques, parmi différentes modalités d'intervention, pour atteindre les ODD et pour produire des arguments puissants en faveur de politiques de changement social et de transition écologique basées sur des preuves empiriques.

Une dernière question pour tous les deux : comment des institutions comme l'AFD et l'IRD peuvent-elles coordonner leurs efforts pour que les pays du Sud bénéficient d'actions de recherche pertinentes qui soient utiles et utilisées pour l'élaboration et la mise en œuvre des politiques ?

Rémy Rioux : Nous pensons que tout le monde gagne lorsque la recherche est menée en partenariat avec les décideurs et la société civile. Ces partenariats doivent préserver l'indépendance et la rigueur de la recherche et favoriser une fertilisation mutuelle afin que la production intellectuelle soit pertinente, appropriée, et qu'elle contribue au progrès de nos sociétés. Cette implication doit se manifester à toutes les étapes du cycle de la production scientifique, depuis la conception et le cadrage de la recherche, à sa mise en œuvre et à la formalisation des connaissances et leur diffusion.

Le groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) est un bon exemple de cette dynamique. Il est essentiel de combiner l'action contre les changements climatiques à des politiques visant à réduire les inégalités et à renforcer le lien social dans nos sociétés pour garantir que la transition environnementale, si vitale aujourd'hui, soit également socialement durable. Cette recherche dont nous avons besoin devrait être implantée dans le Sud, avec le soutien et l'accompagnement de centres d'excellence du Nord lorsque c'est pertinent. « Le Sud inspirera le Nord », pour citer le président d'Unicef France, Jean-Marie Dru (communication personnelle).

À cet égard, l'IRD est un partenaire précieux, car il est reconnu pour son excellence scientifique et ses liens étroits avec les équipes universitaires des pays du Sud. Sa mission est en outre entièrement dédiée aux pays en développement, et tous ses travaux sont menés en partenariat et pour renforcer les capacités de recherche dans le Sud. Nos deux institutions sont engagées dans une démarche conjointe depuis 2012 en matière de recherche sur le développement durable. Notre collaboration s'appuie sur l'approche de la science de la durabilité, promouvant l'interdisciplinarité et le rapprochement entre connaissances scientifiques et les savoirs des autres acteurs du développement.

Dans la même veine, l'International Development Finance Club (IDFC), que je préside depuis 2017, œuvre pour que les pays du Sud bénéficient de travaux de recherche pour élaborer et mettre en œuvre des politiques publiques. Le Club a, de fait, présenté récemment un rapport novateur, rédigé par les groupes de réflexion indépendants CPI (Climate Policy Initiative) et I4CE (Institute for Climate Economics), qui fournit un cadre solide, utilisable par les 26 banques de développement nationales et régionales membres de l'IDFC – dont beaucoup sont basées dans les pays du Sud –, et par la communauté financière dans son ensemble, afin d'aligner les stratégies de toutes les institutions financières avec les objectifs de l'accord de Paris aux niveaux national, stratégique et opérationnel. L'évaluation des projets fait également partie des sujets discutés au sein du Club afin d'identifier les questions soulevées par l'évaluation des actions pour le climat par les organisations internationales et les bailleurs de fonds, et notamment l'examen des défis méthodologiques liés à la mesure des impacts des programmes de développement concernant le changement climatique. Dans les mois à venir, le lancement de la plateforme électronique de l'IDFC permettra au Club de favoriser les échanges entre les experts et de promouvoir le partage des connaissances et des bonnes pratiques.

Jean-Paul Moatti : L'IRD a derrière lui une longue histoire, puisqu'il a fêté ses 75 ans en 2019. Il travaille dans plus de 50 pays en développement. L'AFD bénéficie de l'expertise des chercheurs de l'IRD depuis de nombreuses années. Paradoxalement toutefois, bien que la France reste le seul pays avancé à disposer, avec l'IRD, d'un organisme public de recherche interdisciplinaire, dont la mission unique est la coopération scientifique équitable avec les collègues, chercheurs et universitaires, des pays en développement, nous avons travaillé moins fréquemment et moins systématiquement avec l'AFD que d'autres agences et banques de développement – comme l'USAID et le DFID britannique – ne l'ont fait avec leurs chercheurs nationaux. L'une des raisons est que, jusqu'à une date récente, le financement de la recherche ne faisait pas partie de la mission de l'AFD, les contributions scientifiques devant donc emprunter des canaux d'expertise contractuelle qui n'étaient pas toujours bien adaptés aux projets de recherche. Mais la situation évolue rapidement, alors que les ODD et l'action contre le changement climatique deviennent des axes de travail communs à l'AFD et à l'IRD, et que le rôle majeur de la diplomatie scientifique dans le développement durable est de plus en plus reconnu (voir le GIEC pour le climat et l'Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services [IPBES] pour la biodiversité et, bien sûr, le GSDR).

Premièrement, et avant tout, l'AFD soutient maintenant des programmes de renforcement des capacités universitaires et de recherche dans les pays en développement, comme on peut le voir en Côte d'Ivoire, où l'AFD intervient, conjointement avec le programme français de désendettement et dans sa collaboration avec la Banque mondiale, pour soutenir les Centres d'excellence africains (African Center of Excellence – ACE) dans les domaines de la recherche, notamment en Afrique francophone. De nombreux partenaires africains du projet ACE de l'IRD reçoivent aujourd'hui l'aide de l'AFD. Deuxièmement, le nouvel

accord global signé début 2019 entre l'AFD et l'IRD vise à associer plus étroitement, dans les pays, les initiatives décentralisées menées par l'AFD avec des chercheurs aux grands programmes codéveloppés par les deux organisations. Ces programmes incluront des projets d'évaluation scientifique et de développement financés par l'AFD, qui pourraient s'avérer riches d'enseignements et permettre un certain niveau de généralisation à l'échelle internationale. L'élaboration d'un protocole d'évaluation expérimental ou quasi expérimental *ex ante*, randomisé ou non, et la mise en place simultanée d'un projet de développement durable augmentent les chances de réussite, tant pour l'évaluation scientifique que pour le projet lui-même. Une collaboration plus étroite avec l'AFD est un excellent moyen de promouvoir les changements requis par la science de la durabilité en matière de pratiques de recherche, ceci incluant la coconstruction de programmes de recherche avec les communautés et les groupes de population vulnérables qui sont directement concernés.

Entretien avec Gulzar Natarajan

Gulzar Natarajan, au cours de votre carrière de haut fonctionnaire du gouvernement indien, vous avez accumulé un grand nombre d'expériences. Vous avez occupé un poste au sein du cabinet du Premier ministre, vous avez dirigé la société de développement des infrastructures de l'État d'Andhra Pradesh, vous avez été collecteur de district[1] *à Hyderabad, vous avez été président-directeur général d'une société de distribution d'électricité basée à Visakhapatnam, vous avez été commissaire municipal de Vijayawada, et vous avez occupé des postes dans le domaine du développement dans l'État d'Andhra Pradesh. Vous avez dirigé la conception et la mise en œuvre de projets de grande envergure dans de nombreux secteurs, tels que les infrastructures, l'urbanisme, la santé, l'éducation, le renforcement des moyens de subsistance ou encore la réduction de la pauvreté.*

Ces différentes fonctions, combinées à votre formation en ingénierie d'une part et en études du développement d'autre part, vous permettent d'avoir une vision très précise du fonctionnement des bureaucraties indiennes et des politiques de développement, de la façon dont elles peuvent être améliorées et du type de méthodologie de recherche susceptible de contribuer à cette amélioration.

Quelles sont, selon vous, les questions les plus importantes en matière de politiques publiques auxquelles l'Inde est confrontée aujourd'hui et qui pourraient être éclairées par des recherches solides, en commençant par le niveau macro ?

Il existe un certain nombre de problématiques majeures auxquelles les décideurs politiques indiens sont confrontés de façon régulière. Elles pourraient toutes tirer profit d'informations et de preuves empiriques issues de recherches de haute qualité. J'ai moi-même été confronté à plusieurs d'entre elles à un moment ou à un autre, et j'ai été frustré par le manque de recherches qui auraient pu éclairer davantage mes hypothèses de travail et les solutions proposées. J'ai recensé douze domaines que je considère comme essentiels : la macro-économie, les marchés financiers, les infrastructures, le secteur bancaire, la politique industrielle, les finances publiques, les réformes du marché du travail, les marchés informels, l'urbanisation, le développement, la politique étrangère et commerciale et l'analyse des données. J'énumère dans l'annexe A les sous-questions spécifiques qui me paraissent fondamentales dans chacun de ces domaines.

1. Note sur la traduction : En Inde, le collecteur est le représentant de l'État au niveau du district. Il est chargé de superviser la collecte d'impôts, de certaines fonctions de maintien de l'ordre et de la gestion de crise.

Outre ces grands sujets, pourriez-vous nous donner quelques exemples concrets pertinents axés sur des questions plus pratiques qui relèvent de la compétence des maires ou des commissaires municipaux ?

En Inde, il existe trois niveaux de gouvernement à l'échelle infranationale : le gouvernement local (urbain ou rural), l'administration de district et les services des gouvernements des États. Les bureaucrates et les dirigeants politiques à ces niveaux sont confrontés à de multiples défis et doivent prendre des décisions en temps réel sur la base de données et d'informations limitées.

Quelles sont les questions, à la fois stratégiques et opérationnelles, qui occupent généralement l'esprit des dirigeants à chacun de ces trois niveaux ? Quels sont les choix de solutions techniquement fiables, administrativement réalisables et politiquement acceptables pour chaque problématique ? Comment une recherche scientifique de haute qualité peut-elle informer et nourrir les bureaucrates dans la prise de décision et dans la mise en œuvre ? Quelles sont les méthodes de recherche appropriées pour soutenir ce processus ?

Pour donner une idée des défis à relever, permettez-moi d'exposer les différentes questions auxquelles est confronté un bureaucrate type à chacun des trois niveaux de gouvernement dans n'importe quel État indien. Sans viser l'exhaustivité, mon objectif est de recenser un maximum de grands domaines d'intervention possible à chaque niveau. Compte tenu de la taille du pays, ces niveaux administratifs couvrent pour certains au moins un million de personnes, et même des populations qui se comptent en dizaines de millions au niveau des États. En termes d'impact également, chacune de ces questions s'étend sur l'ensemble du système particulier, et un changement de la pratique habituelle peut avoir des effets importants, souvent transformationnels. Il s'agit donc bien de défis de développement de premier ordre ayant un impact majeur.

En outre, ces questions définissent les problématiques communes et constituent un bon point de départ pour la recherche, représentant ainsi une bonne pratique pour s'attaquer à un sujet difficile.

L'annexe B fournit la liste des questions qui mobilisent les réflexions des responsables politiques à chacun de ces trois niveaux.

Il est évident que l'exploration des réponses à bon nombre de ces questions ne dépend pas d'une méthode de recherche particulière. En effet, la majorité d'entre elles ne relèvent peut-être pas d'approches quantitatives pures et nécessiteraient une analyse qualitative et ethnographique. Un consultant équipé d'une panoplie d'outils suffisamment rigoureux pourrait s'avérer le mieux placé pour les explorer. D'autres cas peuvent nécessiter une combinaison de techniques allant de l'analyse des données aux méthodes économétriques et aux expérimentations de terrain.

Comme on peut le constater, parmi ces questions, seul un petit nombre se prête à une RCT rigoureuse. Pour commencer, dans la plupart des cas, le bureaucrate n'a pas la flexibilité requise pour créer un traitement et des contrôles. Deuxièmement, il est le plus souvent impossible d'isoler clairement

un problème et ses solutions potentielles afin de pouvoir examiner l'attribution. Troisièmement, comme les interventions portent sur des systèmes vastes et complexes, où les facteurs influents ne sont pas simples à identifier, il est plus facile d'étudier les contributions que les attributions. Quatrièmement, les bureaucrates prennent des décisions en temps réel et ne peuvent donc pas se permettre de longues expérimentations. Cinquièmement, les résultats immédiats de ces interventions ne reflètent bien souvent que des équilibres partiels, et il faut beaucoup de temps avant de pouvoir observer des résultats stables. C'est pour cela que les évaluations globales atteignent rarement leur objectif. Enfin, en rapport avec le point précédent, il n'existe pas de solution unique qui soit susceptible d'être simplement reprise et mise en œuvre. Au lieu de cela, les problèmes (y compris le sabotage interne) commencent à émerger lorsque la solution est lancée, ce qui nécessite une adaptation itérative, surtout dans la période initiale, avant la stabilisation.

En ce qui concerne les expérimentations de terrain, le test A/B, simple et rapide, peut apporter des informations sur les choix concurrents évidents à des carrefours décisionnels pendant la mise en œuvre, et s'avérer une technique plus pertinente que les RCT de longue durée. L'idée est de vérifier si certains indicateurs immédiats de succès probable (identifiés à partir d'un examen de la théorie du changement pour l'intervention) sont satisfaits. Face à des questions aussi complexes et à des environnements difficiles, et lorsque les résultats mettent du temps à se manifester, il peut être plus pertinent et plus utile de vérifier le respect d'indicateurs de processus immédiats plutôt que de procéder à des évaluations globales de ces résultats.

Concernant les questions soulevées par les partisans des RCT, apportent-elles les bonnes réponses, et dans quelle mesure ces réponses peuvent-elles être utiles ?

Je répondrai par trois exemples précis. Prenons le cas de la lutte contre l'alcool au volant, qui a été étudiée par Abhijit Banerjee et ses collègues (Banerjee *et al.*, 2012) et résumée dans une publication (Banerjee *et al.*, 2017c). Ils plaident en faveur d'une utilisation accrue des éthylomètres (outils) et de l'introduction d'amendes plus élevées (lois) pour lutter contre la conduite en état d'ivresse. Ils concluent également que ce dispositif doit être complété par une stratégie de contrôle des véhicules dans des lieux définis de manière aléatoire, en faisant appel à des équipes de police dédiées, issues des forces de réserve.

Cette « stratégie » n'a rien de nouveau. Elle est couramment appliquée par les commissaires de police sur de courtes périodes lorsque quelque chose (généralement un accident très médiatisé ou une directive du tribunal) suscite une vigilance accrue en matière de conduite en état d'ivresse. Le problème est que ces dispositifs ne peuvent pas être étendus au-delà de courtes périodes.

Le défi, comme l'a souligné Esther Duflo dans d'autres contextes, est une question de « plomberie ». Prenons la question des lieux aléatoires. Si, vue de l'extérieur, l'allocation aléatoire de lieux peut ressembler à un exercice

algorithmique, il peut être difficile de la rendre opérationnelle à grande échelle. Face à des intérêts et des systèmes puissants qui incitent à « serrer les rangs », la capacité d'action institutionnelle est faible, en particulier dans les commissariats de police. De ce fait, les stratégies comme la vigilance accrue sont facilement compromises ou affaiblies, sauf si des dirigeants extrêmement engagés assurent eux-mêmes la gestion du processus.

À titre d'exemple, les contrôles qualité aléatoires effectués par des tiers sur des travaux d'ingénierie en cours, qui sont aujourd'hui monnaie courante, sont fréquemment compromis par la collusion. En ce qui concerne les réservistes, ils ne peuvent pas être mobilisés au-delà de quelques jours. Avec des forces de police en sous-effectif flagrant et débordées par les responsabilités de gestion de foule et des événements VIP qui exigent des services de sécurité, ils n'ont de « réserves » que le nom, tellement les demandes concurrentes sont nombreuses.

Par ailleurs, les normes en matière de preuve pour établir une infraction légale de conduite en état d'ivresse peuvent restreindre le rôle des réservistes. Par exemple, afin de limiter l'excès de pouvoir discrétionnaire, la loi (c'est-à-dire les règlements adoptés en vertu de la loi centrale dans différents États) impose que les tests d'alcoolémie soient effectués en présence d'un représentant de police d'un certain rang. Or, les administrations de police, débordées, disposent de trop peu de personnels de ce rang pour les dédier à la circulation, et encore moins aux patrouilles de nuit traquant les cas de conduite en état d'ivresse. Mais la délégation de cette responsabilité pose des problèmes juridiques et pratiques.

Les conclusions ci-dessus découlent d'une RCT réalisée dans 162 commissariats de police du Rajasthan et couvrant cinq interventions en matière de gestion : la limitation des mutations arbitraires, la rotation des missions et des jours de congés, l'engagement accru de la communauté, la formation interne et les visites « secrètes » par des agents de terrain se faisant passer pour des civils. L'étude a montré que les trois premières interventions « qui auraient réduit l'autonomie des cadres intermédiaires, étaient mal mises en œuvre et inefficaces », alors que les deux dernières avaient « des impacts solides ». Sur la base de ces observations, les chercheurs ont pointé les « résultats très positifs » d'une intervention associant une « bonne performance » en matière de tests de sobriété *sans s'appuyer sur les cadres intermédiaires,* à la « promesse d'une mutation des casernes de réservistes à un poste attractif dans un commissariat de police ».

La publication allègue que « les résultats expérimentaux exposés dans cet article montrent qu'il est possible d'influencer le comportement de la police dans un laps de temps relativement court, en utilisant un ensemble d'interventions simples et abordables ». Cette affirmation est trompeuse.

Examinons les « stratégies » explorées par les chercheurs. Nous avons déjà parlé du défi que représentent les contrôles aléatoires et le recours aux réservistes. Les « formations internes » sont un élément incontournable pour tout système administratif. Le seul problème est que les formations ne se traduisent pas par un apprentissage ou une appropriation suffisamment utiles. C'est d'ailleurs un

défaut courant des formations internes dont bénéficient les enseignants, les médecins, les inspecteurs, etc.

Les visites « secrètes » par des agents de terrain ne constituent pas une idée nouvelle. Il s'agit d'un élément de base du renseignement d'origine humaine. Là encore, il existe des difficultés pratiques pour identifier et gérer ces activités. En outre, ces visites peuvent également créer des distorsions systémiques qui font plus de mal que de bien. Ainsi, plutôt que de se focaliser sur de telles mesures fantaisistes et sporadiques, les chefs de police devraient concentrer leurs efforts sur l'amélioration de leurs services de renseignement et de leurs services spéciaux, et sur l'utilisation de multiples canaux incluant les visites secrètes, les organismes tiers et la sollicitation de retours téléphoniques de la part des plaignants. En fait, de nombreux chefs de police en Inde varient régulièrement ces différentes approches pour obtenir des retours d'information.

Prenons le cas de la dernière intervention, qui associe les mutations de réservistes aux bonnes performances et réduit l'autonomie des cadres intermédiaires. Là encore, le défi relève de la « plomberie ». Pour commencer, la lutte contre la conduite en état d'ivresse ne figure bien souvent pas parmi les priorités des forces de police classiques, si l'on met de côté les forces de réserve. Et les activités principales des réservistes, c'est-à-dire les services de sécurité, ne se prêtent pas à des évaluations quantitatives individuelles des policiers. Deuxièmement, dans quelle mesure une politique qui cherche explicitement à récompenser certains policiers par des mutations dans des postes « attractifs » (en échange d'une « bonne performance », quelle qu'elle soit) et à en pénaliser d'autres (le corollaire naturel) en les affectant aux forces de réserve est-elle durable ?

Troisièmement, à partir du moment où l'on commence à associer des mesures incitatives à des indicateurs de performance quantitatifs dans la détection des cas de conduite en état d'ivresse, on n'est pas bien loin d'une fixation d'objectifs qui nous mène vers une pente dangereuse. C'est une caractéristique commune des systèmes publics qui cherchent à associer des mesures de performance aux décisions importantes en matière de gestion du personnel. Quatrièmement, quelle est la durabilité d'un processus administratif où il n'y a aucune implication des cadres intermédiaires ? En fin de compte, il faut bien une gestion institutionnelle du dispositif par des responsables d'un niveau ou d'un autre. Et à supposer même que cette gestion soit assurée par des responsables autres que les « cadres intermédiaires », devons-nous pour autant moins nous soucier d'eux ? Et est-ce même envisageable de penser que ces activités administratives peuvent se faire sans l'implication des cadres intermédiaires ? Enfin, il est faux de prétendre que « la direction de la police ne dispose généralement pas de preuves empiriques et d'informations générées par l'évaluation ». Les dirigeants de la police, pour autant qu'ils gardent les yeux et les oreilles grands ouverts, ont accès dans une large mesure à des « preuves empiriques » d'une portée bien plus vaste, et d'un niveau de crédibilité et de rigueur largement suffisant. Ce n'est pas la production de preuves quantitatives minutieuses qui va vous permettre d'aller bien loin dans les efforts déployés pour gérer efficacement des systèmes de grande ampleur.

Les chercheurs soulignent que les commissions successives de réforme de la police, non seulement n'ont pas défendu les trois interventions « réussies », mais ont également recommandé les interventions « inefficaces ». Pour commencer, les commissions de réforme préconisent des réformes institutionnelles telles que la limitation des mutations, la participation de la communauté, etc. et préfèrent éviter les mesures opérationnelles ordinaires et banales, comme les visites secrètes ou les formations internes, et encore moins celles qui ne sont pas réalisables, comme la gestion des performances des réservistes. En outre, ces recommandations sont des interventions de « plomberie » essentielles pour tout système administratif bien gouverné. En revanche, les solutions proposées par les chercheurs, comme nous l'avons vu plus haut, souffrent de graves lacunes sur le plan pratique. Les instigateurs des commissions de réforme se sont abstenus de recommander de telles solutions de fortune : étant eux-mêmes des « plombiers » de longue date, ayant dû faire face avec plus ou moins de réussite (ou d'échecs) aux défis de « plomberie » qu'implique le maintien de l'ordre dans le monde réel, ils ont été suffisamment responsables et honnêtes pour ne pas le faire.

Ce que je veux dire, c'est que ces idées – les contrôles aléatoires pour la lutte contre la conduite en état d'ivresse, la limitation des mutations arbitraires, la rotation des missions et des jours de congé pour le personnel de police, la participation accrue de la communauté, la formation interne et les visites secrètes par des agents de terrain se faisant passer pour des civils – sont toutes bonnes et n'ont pas besoin de preuves. Ce n'est pas un manque de preuves qui empêche leur adoption. Toutefois, leur mise en œuvre à grande échelle est difficile, et dépend de l'intérêt et de l'engagement du chef de police concerné, et certaines d'entre elles nécessitent des ressources et des capacités que le système ne possède pas actuellement. Dans les systèmes où les capacités de l'État sont faibles, de telles interventions reposent sur des individus (les chefs) et non sur les institutions.

Permettez-moi de prendre un autre exemple : il s'agit d'une RCT concernant des audits menés par des tiers d'installations industrielles polluantes au Gujarat, qui a fait l'objet d'un article (Duflo *et al.*, 2013) et d'un document d'orientation (J-PAL, 2013). Les auteurs affirment apporter la preuve que les audits effectués par un organisme tiers indépendant sont efficaces pour réduire la pollution environnementale.

Pour résumer, en réponse à une directive de la Haute Cour, certaines usines très polluantes du Gujarat commandaient et présentaient des rapports d'audit établis par des organismes tiers trois fois par an depuis 1996. Mais leurs performances étaient peu satisfaisantes. Les chercheurs ont montré qu'une fois que les auditeurs, au lieu d'être payés par les entreprises elles-mêmes, ont été rémunérés par le biais d'un fonds central, que les audits ont été réalisés de manière aléatoire et que les auditeurs ont reçu une prime à la précision, il y a eu une augmentation significative de la présentation des relevés de pollution, et une réduction de la pollution réelle elle-même. Afin de renforcer leur théorie du changement, les chercheurs ont également annoncé et mené des contre-audits de vérification

d'un échantillon aléatoire des relevés de chaque auditeur, et la rémunération des auditeurs a été subordonnée à l'exactitude de leurs contrôles originaux (qui ont été comparés aux contre-audits de vérification).

Aujourd'hui, personne ne conteste la valeur des audits réalisés par des agences indépendantes payées à partir d'un fonds central, et renforcés par des contre-audits de vérification. Les audits ou certifications réalisés par des tiers sur un échantillon aléatoire (et de manière inopinée) sont aujourd'hui monnaie courante pour contrôler tout, des travaux d'ingénierie à la qualité des biens achetés et des services fournis. En Inde, au cours des deux dernières décennies, les audits de qualité menés par des tiers, manifestement effectués sur des échantillons aléatoires et de manière inopinée, ont été adoptés pour les travaux d'ingénierie exécutés par toutes les structures, qu'elles soient petites ou grandes, urbaines ou rurales. Ils ont sans aucun doute contribué à améliorer la qualité de ces travaux et, lorsqu'ils sont bien réalisés, leurs bénéfices sont très importants. Et les responsables de la protection de l'environnement les plus avisés, présents dans tous les États, sont bien conscients de leur utilité.

Peut-on ainsi dire que la recherche a apporté des informations précieuses aux bureaux de contrôle de la Pollution (*Pollution Control Boards* – PCB) en Inde ? Dans les contextes où la capacité de l'État est faible, la gestion efficace des audits menés par des tiers représente elle-même une tâche énorme. Les contre-audits de vérification rendent la démarche encore plus lourde. Étant donné que ces contre-vérifications ont été effectuées sous la supervision de chercheurs de l'équipe, enthousiastes et impliqués, au su des entreprises, et que la rémunération des auditeurs était subordonnée à la concordance entre les audits initiaux et les contre-audits, les audits initiaux ont forcément été assortis d'enjeux plus importants, et sont donc aussi devenus plus qualitatifs (effet Hawthorne). La RCT a démontré l'efficacité de cette méthode particulière de double audit.

De tels audits, combinant deux niveaux, ambition et qualité, sont bien évidemment souhaitables. Malheureusement, ils sont bien trop exigeants et lourds pour avoir une chance d'être généralisés de façon efficace dans des systèmes publics de faible capacité, *a fortiori* lorsque la pollution est la norme plutôt que l'exception.

Dans l'intervention précédente, les chercheurs ont observé au moins deux sources de distorsion par les incitations. Premièrement, un problème d'agencéité découlant du fait que les auditeurs étaient payés par les structures auditées. Deuxièmement, le fait que les auditeurs recevaient une rémunération nettement inférieure à celle qui aurait été requise pour réaliser de bons audits.

Concernant le problème d'agencéité lié à la façon dont les auditeurs sont rémunérés, les chercheurs n'avaient pas besoin de recourir à une RCT, car il existe une littérature très riche sur les problèmes liés à l'achat des notations de crédit par les institutions financières. L'idée que les auditeurs/agences de notation ne soient pas rémunérés par les structures auditées/notées est largement acceptée. Ensuite, les auditeurs sont de moins en moins bien rémunérés, car ces audits sont devenus en grande partie un exercice pour la forme, et toutes les parties le savent.

Il est difficile de croire que le PCB du Gujarat ne savait pas ce qui se passait. Je vois au moins cinq bonnes raisons relevant de ladite « plomberie » pour lesquelles le PCB a préféré s'en tenir au *statu quo* plutôt que d'adopter des réformes qui paraissaient évidentes (ne serait-ce que parce que, comme mentionné précédemment, elles étaient déjà en cours dans d'autres secteurs). Premièrement, ces rapports étaient produits à la demande de la Haute Cour et lui étaient transmis. Tant que la Cour était satisfaite, il n'y avait aucune raison ni motivation intrinsèque pour que le PCB change de système. Les cas de conformité pour la forme aux exigences réglementaires ne sont pas rares. Deuxièmement, si le gouvernement décidait d'effectuer les audits, la question se posait alors de savoir qui allait les payer ou comment les montants seraient perçus, sans parler du casse-tête que représentait la gestion de cette responsabilité administrative supplémentaire. Troisièmement, les intérêts particuliers au sein des industries polluantes étaient forts et le *statu quo* préférable. Dans de tels contextes en particulier, la capture réglementaire et la tolérance administrative ne sont jamais bien loin. Pour ces entreprises en effet, la différence entre le *statu quo* et le respect de la réglementation en matière de pollution est, dans de nombreux cas, une question de survie, avec des répercussions importantes sur l'emploi. Quatrièmement, un durcissement brutal et rigoureux des normes entraînant un changement radical au niveau des rejets d'effluents aurait pour effet la fermeture de plusieurs usines, et, comme l'ont montré de nombreux précédents, cela perturbe des économies politiques fragiles. Enfin, que cela nous plaise ou non, la lutte contre la pollution a longtemps été une préoccupation marginale pour la plupart des gouvernements des États, qui courent après la croissance économique et la création d'emplois. Leur volonté d'entreprendre des réformes rigoureuses s'est par conséquent avérée limitée.

Prenons un dernier exemple. Karthik Muralidharan et ses collègues ont réalisé une RCT sur l'utilisation des téléphones mobiles pour améliorer la gouvernance (MURALIDHARAN *et al.*, 2018a ; 2018b). Pour résumer, un retour d'information basé sur des appels téléphoniques a été sollicité concernant la qualité de la mise en œuvre du programme Rythu Bandu par le gouvernement de l'État de Telangana, dans le cadre duquel des transferts monétaires directs par chèque ont été effectués aux agriculteurs éligibles (les agriculteurs ont-ils bien reçu le chèque, l'ont-ils reçu à temps, l'ont-ils encaissé, etc.). Une évaluation des appels téléphoniques au moyen d'une RCT a révélé que 83 % des agriculteurs avaient reçu et encaissé leur chèque, que les agriculteurs situés dans les zones faisant l'objet de ce suivi avaient 1,5 point de pourcentage de probabilités supplémentaires de recevoir et d'encaisser leur chèque, et 3,3 points de plus dans le dernier quartile des fermiers propriétaires. Selon les auteurs, les centres d'appels ont distribué 70 millions de roupies supplémentaires (soit 1 million de dollars US) aux agriculteurs, et ont coûté 2,5 millions de roupies (soit environ 35 000 dollars US).

Il convient à nouveau de s'interroger : qui conteste l'efficacité de cette idée ? Cela nécessitait-il une RCT ? La pratique consistant à solliciter un retour d'information des citoyens et des clients par le biais d'appels téléphoniques est

en vogue depuis plusieurs années. Plusieurs organismes publics à travers les États ont mis en place depuis des années des systèmes de sélection des citoyens pour ces retours d'informations. L'État voisin d'Andhra Pradesh a même fait de cette pratique un élément central des études de performance en évaluant la plupart des activités gouvernementales par le biais de remontées d'information par téléphone réalisées grâce à son système élaboré de gouvernance en temps réel (*Real Time Governance System* – RTGS). Depuis longtemps, plusieurs compagnies de distribution d'électricité et de corporations municipales disposent de centres d'appels téléphoniques de ce type, leur permettant d'obtenir des retours d'informations.

Aucun bureaucrate ne viendrait contester l'idée sous-jacente que le retour d'information par le biais d'appels téléphoniques de citoyens échantillonnés de manière aléatoire constitue un moyen utile d'évaluer la qualité de la mise en œuvre. Cela soulève deux questions. Premièrement, s'agit-il là de l'approche la plus durable et rentable pour améliorer la qualité de la mise en œuvre ? Par exemple, comme nous le verrons plus loin, l'amélioration du système de suivi existant aurait été possible sans coût supplémentaire et avec d'autres avantages.

Deuxièmement, même si nous poursuivons le retour d'information par téléphone, qu'en est-il des problèmes qui en découlent ? D'abord, avec un système de retour d'information par téléphone, le véritable défi n'est pas d'obtenir un retour d'information, même granulaire et exploitable, mais la capacité du système à réagir à ces retours d'information d'une manière suffisamment constructive. La véritable limite est là, et elle dépend essentiellement de la capacité de l'État à traiter activement une problématique élémentaire de gouvernance : contrôler l'information et agir efficacement sur la base de celle-ci. En outre, nous ne devrions pas écarter le scénario d'un effet d'échelle dans lequel ce contrôle, en l'absence d'exigences de suivi, a toutes les chances de venir s'ajouter à la panoplie d'outils de surveillance sans apporter le moindre avantage supplémentaire, mais avec un coût additionnel important. De plus, la gestion des centres d'appels téléphoniques et du système de gestion des retours d'information lui-même monopolise des ressources administratives par ailleurs limitées. Enfin, ce système peut fausser les dispositifs d'incitation au sein de la bureaucratie et affaiblir les mécanismes de contrôle existants.

En fait, on pourrait très facilement imaginer le scénario suivant : les systèmes de retour d'information par téléphone peuvent améliorer l'efficacité de la mise en œuvre, et ils ont un petit air nouveau et séduisant. Alors, mettons en place des centres d'appels téléphoniques dans chaque district/État. Et dans cinq ans, ces centres d'appels dysfonctionnels et les énormes ressources financières qu'ils auront mobilisées viendront s'ajouter à la liste funeste des innovations de développement sans lendemain.

Soit dit en passant, j'ai participé en personne à la mise en œuvre des trois interventions exposées dans les publications citées ci-dessus, et j'ai été confronté à toute la confusion et autres défis pratiques engendrés par leur application.

Audits menés par des tiers et remontées d'informations par téléphone ont été une constante dans les différents postes que j'ai occupés depuis 2005 au moins.

Toujours à propos des RCT, quelles seraient vos propres réponses et où se situent-elles par rapport aux suggestions des évaluations randomisées (par exemple dans le domaine de la gouvernance, pour améliorer les performances de l'administration indienne) ?

Commençons par le cas de la police. Au lieu de proposer des solutions qui contribuent à améliorer les institutions et les systèmes dans lesquels s'inscrivent ces activités de maintien de l'ordre, les chercheurs en viennent à recommander des solutions ponctuelles et non durables qui ne tiennent pas suffisamment compte des facteurs contextuels et des considérations systémiques.

Concernant la lutte contre la conduite en état d'ivresse, la recherche aurait par exemple dû avoir pour objectif d'améliorer les résultats de la police en *renforçant la responsabilisation des cadres intermédiaires* et en *trouvant des solutions institutionnelles durables*, au lieu de *se passer des cadres intermédiaires et de faire appel ponctuellement à des visites secrètes*.

Les incitations à la performance qui impliquent des récompenses financières ou des mutations d'une ampleur suffisamment grande ont peu de chances de fonctionner à grande échelle dans des systèmes publics complexes. En fin de compte, où ces mesures ont-elles une chance de fonctionner, même dans les pays développés ? Pour commencer, il est extrêmement problématique de quantifier les résultats de manière crédible, et encore plus de les collecter et les gérer, dans la plupart des contextes considérés. Des preuves de réussite limitées, constatées pour des activités logistiques contrôlables, comme le recouvrement des impôts, ne signifient pas que l'on peut appliquer les mêmes mesures avec des attentes similaires pour les enseignants ou la police. Deuxièmement, dans la durée, il y a une grande probabilité que les incitations financières finissent par être considérées comme un dû, aggravant ainsi le problème des salaires déjà élevés des fonctionnaires de niveau inférieur. Enfin, les affectations et les mutations font partie des actions administratives dont les enjeux sont les plus importants, et lorsqu'elles interviennent dans des contextes où les « bonnes performances » ne peuvent pas être hiérarchisées de manière crédible et incontestable, elles peuvent entraîner polémiques et mécontentement.

Il existe très peu d'innovations, qu'il s'agisse de grandes idées, de refontes des processus, voire de nouvelles théories de management, qui puissent durablement et significativement « influencer le comportement de la police dans un laps de temps relativement court » dans des contextes où le système et sa gouvernance sont entachés de graves lacunes. Toutes choses égales par ailleurs, les systèmes de police qui « fonctionnent bien » reposent vraisemblablement sur la combinaison d'une capacité administrative fonctionnelle et d'un bon leadership. L'intensité du second facteur peut même masquer temporairement les déficiences du premier. C'est pour cette raison que nous continuons à entendre parler d'administrations mal gérées qui deviennent soudainement

efficaces à l'arrivée d'un bon commissaire de police, et reviennent à l'état précédent lorsque celui-ci s'en va.

En langage économique, la fonction de production permettant d'obtenir de bons résultats au niveau de la police dépend essentiellement de ces deux éléments. La plupart du temps, les innovations peuvent fonctionner à la marge pour améliorer la capacité administrative et libérer des énergies de leadership pour une utilisation productive à un autre niveau. Mais dans les systèmes vraiment faibles, comme c'est le cas ici, le leadership est nécessaire afin de générer de bons résultats à court terme et renforcer les capacités institutionnelles à long terme.

Relever le défi du renforcement des capacités de l'État était une opportunité offerte aux chercheurs engagés dans le programme Rythu Bandhu au Telangana. Pour les chercheurs, le problème de fond était d'améliorer l'efficacité de la mise en œuvre du programme. Le système de remontées d'informations par téléphone n'est qu'une approche parmi d'autres, qui présente plusieurs dangers potentiels. Les chercheurs auraient plutôt pu profiter de l'occasion pour comprendre comment le suivi de la mise en œuvre pouvait être amélioré. Cela aurait été utile pour étudier les problèmes de capacité de l'État, et notamment la question cruciale d'une utilisation plus efficace des systèmes de surveillance permettant de contrôler la qualité de la mise en œuvre du programme. Ces problématiques ne se prêtent malheureusement pas aux RCT.

Que serait une approche de suivi plus durable et plus efficace pour étudier les programmes de développement ? Considérons certaines variables (il pourrait y en avoir d'autres) liées à des études de ce type : qui étudie et qui est étudié, à quelle fréquence, quels sont les paramètres spécifiques concernés ? Ainsi, un collecteur de district (ou un responsable de développement de bloc, en anglais *Block Development Officer* – BDO) pourrait examiner une fois par semaine (ou une fois tous les quinze jours) le travail du BDO (ou du percepteur de village, en anglais *Village Revenue Officer* – VRO) sur certains indicateurs de processus (traitement des autorisations) ou de résultat (réception ou encaissement).

Comment pourrait-on optimiser cet examen ? Disons que tous les blocs seraient divisés en deux groupes de traitement et deux méthodes d'examen différentes des BDO (ou utilisées par ceux-ci) ou des VRO – on peut imaginer ces options sans trop d'efforts –, que l'on opposerait au groupe de contrôle gardant les méthodes de suivi habituelles.

Un tel dispositif mettrait immédiatement en lumière l'importance de la qualité du suivi et le rôle des améliorations des capacités de l'État (d'une manière plus objective que cela n'a jamais été fait auparavant par des chercheurs) dans une mise en œuvre efficace. Si, par exemple, il s'avère que les BDO qui examinent les VRO une fois par semaine, sont, par rapport à ceux qui procèdent à cet examen une fois par mois, associés à une augmentation de x % des encaissements de transferts monétaires pour le quartile le plus pauvre, cela sert de déclencheur, avec à l'appui des données significatives et exploitables, pour le Secrétaire du gouvernement de l'État de Telangana qui met en œuvre le programme Rythu

Bandhu. On pourrait donc générer des retombées réelles plus importantes et à un coût quasiment nul.

De plus, les gains dépassent la simple amélioration de l'efficacité de l'intervention spécifique. Cette démarche s'appliquerait probablement à la plupart des interventions mises en œuvre dans cette juridiction. Elle mettrait véritablement en évidence la capacité de l'État, et plus précisément la façon dont un meilleur suivi pourrait accroître l'efficacité de la prestation de services publics. Cela peut sembler aller de soi, mais, dans un monde où chacun est à la recherche d'innovations et de manières différentes de faire les choses, ce qui devrait être totalement évident finit souvent par être marginalisé !

Dans la même veine, les questions qui intéressent les praticiens dans le cas des audits menés par des tiers sont plutôt de l'ordre de la « plomberie ». Ce qui aurait été très utile pour les responsables des PCB concernés, c'est la conception d'un système solide d'audits de pollution réalisés par des tiers indépendants. Quelques questions viennent immédiatement à l'esprit. Quelle serait la conception la plus rentable pour des audits menés par des tiers ? Quel nombre, quelle fréquence et quelle portée des inspections seraient les plus rentables ? Que peut-on faire pour limiter le risque de capture réglementaire ? Comment les audits peuvent-ils répondre à des attentes évolutives ? Comment les méthodes d'inspection devraient-elles changer pour pallier la forte probabilité d'adaptation et de manipulation des audits par les usines ? Et, comment les audits peuvent-ils être financés de manière durable ?

Dans le cas de travaux d'ingénierie, la conception la plus rentable serait axée sur le plus petit nombre d'échantillons avec la périodicité la plus longue possible sans compromettre l'effet dissuasif. Concernant la capture, les inspections devraient non seulement être aléatoires, mais également réalisées par du personnel en rotation. Les organismes eux-mêmes devraient être réorganisés régulièrement, ou plusieurs organismes devraient être sollicités. Pour ce qui est des attentes évolutives, il pourrait s'avérer nécessaire de revoir périodiquement les critères d'audit et de les calibrer pour prendre en compte les adaptations. Sur le plan du financement, il pourrait être opportun que les PCB commandent et financent les audits et, peut-être, qu'ils récupèrent une partie des coûts sous la forme de taxes aux usagers (ou par le biais des amendes perçues, bien que cela puisse avoir des effets pervers) qui seraient versées dans un fonds commun. Ce sont-là autant de détails compliqués relatifs à la mise en œuvre, qui relèvent du rôle du bureaucrate.

Mais, en fin de compte, ce dont les praticiens ont besoin, c'est d'une conception des audits de tiers qui soit simple et applicable sur le plan administratif. Ou, plus précisément, il leur faudrait un document d'appel d'offres qui reprenne toutes ces spécifications de conception. Or, les travaux de recherche précédemment mentionnés n'apportent rien de pertinent sur ce plan.

Une autre recommandation, en rapport avec votre deuxième question, est d'élargir considérablement le champ des questions de recherche pour couvrir la liste que je vous ai proposée plus tôt.

Quels seraient, selon vous, les points clés que les chercheurs devraient retenir de ces exemples ?

D'emblée, je tiens à préciser que mon objectif n'est pas de minimiser l'importance des RCT ou des expérimentations de terrain dans ces domaines. Elles ont une valeur incontestable, mais doivent être considérées dans leur véritable perspective. Il y aura toujours des cas où les gouvernements devront faire un choix entre différentes options possibles. Ces expérimentations peuvent renforcer les arguments en faveur de certaines idées convaincantes et contribuer à créer une dynamique pour leur adoption à grande échelle. Et, tout aussi important, elles génèrent des preuves empiriques qui peuvent aider à éviter les mauvaises idées. À tout le moins, elles fournissent au responsable politique hésitant une base lui permettant d'initier des audits indépendants et/ou par échantillonnage aléatoire. Fondamentalement, nous avons besoin d'une boîte à outils complète d'évaluations qualitatives et quantitatives pour nous aider à générer des informations permettant de concevoir et d'améliorer la mise en œuvre des interventions de développement.

Les trois cas présentés ont en commun une particularité importante, caractéristique de nombreuses RCT. Ce sont des solutions techniques autonomes qui, considérées isolément, ont un attrait certain sur le plan logique. Les visites secrètes sont des techniques simples et astucieuses. Les audits indépendants réalisés par des tiers qui sont payés à partir d'un fonds central et validés par des contre-vérifications sont logiquement inattaquables. Les centres d'appels et les remontées d'informations par téléphone donnent une impression d'indépendance et de simplicité. Ces trois exemples semblent nouveaux ou innovants, en ce sens qu'ils ne constituent pas la norme. Ils ont un attrait irrésistible lorsque l'on prend le problème isolément et que l'on considère la triste photographie, constellée d'échecs, des réponses administratives habituelles. Et il est possible de mener des projets pilotes de courte durée dans tous ces domaines, supervisés par des assistants de recherche de grande qualité, et de produire des preuves d'efficacité.

Malheureusement, quand on arrive au point critique de la mise en œuvre à grande échelle, l'ironie veut que toutes ces caractéristiques attrayantes se transforment en défauts. Certains facteurs, sur lesquels on avait fermé les yeux, deviennent des freins. La capacité de l'État à administrer et à contrôler, qui était masquée par la faible ampleur de l'expérimentation de terrain et la présence de chercheurs dynamiques et enthousiastes, est mise à nu. La logique ne résiste pas aux défis pratiques. Le système reprend le dessus.

La plupart des bureaucrates en sont conscients et les plus avertis renoncent à ces solutions de fortune ponctuelles ou les adoptent dans le cadre d'efforts de réforme systémique. Les responsables politiques en quête de popularité ou de réponses rapides ont une préférence pour ces solutions « au coup par coup », qui sont invariablement abandonnées dès qu'ils changent de poste.

Dans les trois cas, les preuves apportées par la recherche ont un rôle important à jouer dans l'amélioration des systèmes de police, dans la conception d'un

système solide d'audits menés par des tiers et dans le suivi rigoureux de la mise en œuvre du programme. La réponse à la problématique du bureaucrate n'était pas une évaluation randomisée globale impliquant des innovations difficiles à mettre en œuvre. Ce qui était nécessaire, c'était plutôt de trouver des solutions institutionnelles plus durables.

Après une étape complète de résolution de problème, la recherche spécifique aurait dû porter sur un questionnement tel que celui-ci : sous réserve de l'efficacité des inspections surprises aléatoires dans la lutte contre la conduite en état d'ivresse, des audits menés par des tiers pour vérifier les rejets d'effluents et des systèmes de surveillance pour renforcer l'efficacité de la mise en œuvre du programme Rythu Bandhu, quelle serait l'approche la plus durable, pratique et rentable ?

Comme nous l'avons dit précédemment, il aurait fallu un ensemble d'outils plus hétérodoxes ou une combinaison de méthodes quantitatives et qualitatives. Cela aurait impliqué des tests A/B de courte durée pour déterminer les éléments incertains, des ethnographies/études qualitatives pour identifier les processus critiques, etc. Au cours de la mise en œuvre, il pourrait également s'avérer nécessaire d'utiliser les RCT pour une évaluation complète de modèles de programmes concurrents.

Il convient en outre de noter que, parmi ces problèmes de « plomberie », beaucoup n'ont rien de spécifique au Gujarat ou même à l'Inde. Les contextes généraux sont les mêmes : faible capacité de l'État, bureaucraties centralisées marquées par un faible niveau de confiance, ressources limitées, fonctionnaires surchargés et environnements de travail difficiles. Les facteurs relevant de l'économie politique apportent un degré de complication supplémentaire. Ils sont universels dans la plupart des pays en développement. Ainsi, de nombreux problèmes de « plomberie » susceptibles d'être testés comportent des caractéristiques généralisables. Ces efforts de recherche constituent donc peut-être des occasions manquées et, plus inquiétant encore, de nombreux chercheurs ne sont peut-être même pas conscients que les défis décrits plus haut sont bien réels.

Ces arguments sont motivés par la solide conviction que je me suis forgée en observant les origines de nombreuses études de ce type, voire en dirigeant des agences gouvernementales dans des endroits où plusieurs expérimentations de ce type (mais pas les trois exposées ici) étaient réalisées, que la recherche de terrain commence rarement par un besoin ou un problème ressenti par les représentants du gouvernement. Le plus souvent, les chercheurs en chef, ceux qui ont conçu la recherche, tombent sur une idée, habituellement dans le cadre d'une intervention indépendante, et sont motivés par un souci d'utilité. Partant de cette motivation, ils travaillent à contre-courant pour élaborer une hypothèse de « solution », obtenir un financement par un donateur, puis ils prennent contact avec un interlocuteur gouvernemental pour lui soumettre une proposition d'évaluation. Il est peu probable que l'interlocuteur gouvernemental ait un problème, ni qu'il ait un grand intérêt dans le résultat. Pour faire contrepoids, les gouvernements engagent tout au long de l'étude des consultants pour résoudre

des problèmes spécifiques. Ces missions représentent un effort réel. Le travail débute par un examen approfondi du contexte du problème et des possibilités de solution émergent (une approche de type résolution de problème est plus complète), même si l'analyse des solutions manque de rigueur et les résultats sont pris au sérieux – qu'ils soient mis en œuvre ou non.

Bien qu'il existe aussi des difficultés pratiques, le principal obstacle pour s'attaquer directement au problème réside dans une mauvaise compréhension, parmi les chercheurs, des véritables défis liés à la « plomberie ». Il semble bien que les meilleurs « plombiers » (si on met de côté les vrais) soient les praticiens eux-mêmes. Les compétences en « plomberie » relèvent davantage de l'expérience vécue que d'un savoir théorique acquis.

Annexe A. Liste des sujets de recherche sur l'économie et la politique étrangère de l'Inde

Il est préoccupant de constater que des sujets importants relatifs à l'Inde ne suscitent qu'un intérêt limité de la part des chercheurs. Le seul domaine qui intéresse les chercheurs renommés au sujet de l'Inde concerne les études sur la pauvreté comportant une évaluation randomisée (RCT) et des conceptions romancées de l'entrepreneuriat et des entreprises sociales visant la « base de la pyramide » (*Bottom of Pyramid* – BoP).

Les activités de recherche peuvent aller des analyses économétriques aux ethnographies, en passant par les études de cas/d'événements. L'objectif devrait être de promouvoir une recherche de la plus haute qualité qui puisse alimenter de manière fiable les débats sur certaines problématiques de l'Inde, et influencer ainsi l'élaboration de politiques dans ces domaines.

Voici une liste indicative des domaines dans lesquels les politiques publiques peuvent être améliorées de manière significative grâce, notamment, à des travaux de recherche de haute qualité.

1. Marchés financiers

   a. Politique monétaire : comment l'approche indienne hétérodoxe de la politique monétaire menée avant et pendant la crise financière mondiale se situe-t-elle par rapport à l'approche orthodoxe ? Évaluation de la transmission de la politique monétaire, quantification, contraintes probables et comparaison avec d'autres pays. Évaluation des causes récentes de l'inflation en Inde. Le ciblage de l'inflation est-il une stratégie de politique monétaire appropriée pour l'Inde, ou l'Inde devrait-elle adopter un ensemble d'outils plus hétérodoxes et, si oui, de quoi devrait être composé cet ensemble ?

   b. Quelle a été l'expérience de l'Inde en matière de gestion des flux de capitaux, notamment par rapport à ses homologues ? Comment l'Inde a-t-elle réussi à atténuer les répercussions des arrêts soudains et des fuites de capitaux ? Quels sont les enseignements tirés des politiques de gestion des taux de change ?

c. Marchés de capitaux : comment les différentes composantes des marchés de capitaux indiens ont-elles évolué dans le temps par rapport à leurs homologues ? Quel est le niveau d'intégration globale des différents segments des marchés financiers nationaux ? Comment la réglementation des marchés de capitaux de l'Inde se situe-t-elle par rapport à celle de ses homologues ?

2. Infrastructures

a. Quelle a été l'expérience de l'Inde en matière de la participation du secteur privé aux projets d'infrastructures par rapport à celle des pays d'Amérique latine et d'Europe ? Évolution des partenariats public-privé (PPP), tendances en matière de dépassements de coûts et de délais, évaluation du rapport qualité-prix et comparateurs du secteur public par rapport aux PPP, problèmes des offres agressives et des financements imprudents, renégociations et leurs cadres, variantes de contrats/concessions, sources de financement, monétisation des actifs, etc.

b. Comment les entreprises d'infrastructures indiennes, leurs pratiques commerciales et leurs stratégies de financement se situent-elles par rapport à celles d'autres grandes économies (notamment européennes) ?

c. Comment se situent les barèmes de coûts et les pratiques/approches en matière de passation de marchés publics d'infrastructures en Inde, en Chine et ailleurs ?

3. Secteur bancaire

a. Quelle est l'expérience de l'Inde en matière de crises périodiques du secteur bancaire et de leur résolution par rapport à celle de la Suède, des États-Unis, de l'Espagne, de l'Irlande, de l'Islande, de l'Italie, etc. ?

b. Quels sont les pratiques de gestion, les contrôles internes et les processus d'évaluation du crédit des banques publiques et des banques privées ? Comment les pratiques de micro-gestion du gouvernement faussent-elles les mécanismes d'incitation au sein des banques du secteur public ?

4. Politique industrielle

a. Quelle est l'évaluation/la quantification des zones économiques spéciales (ZES) de l'Inde sur le plan de la création d'emplois, de l'augmentation de la production et du vaste champ des externalités, y compris les retombées technologiques et les effets de déplacement ? Comment se situent-elles par rapport à la politique de la Chine en matière de ZES ? Comment une approche basée sur des « Villes sous charte » (*Charter Cities*, en anglais), ou sur des réglementations pourrait-elle être différenciante, et quels gains potentiels peuvent être quantifiés ?

b. Quelle est l'évaluation/la quantification de la politique industrielle de l'Inde basée sur les avantages fiscaux et les concessions sur les intrants, à la fois en termes de bénéfices directs et d'externalités ? Quel a été son impact sur la croissance des petites et moyennes entreprises (PME) par rapport aux grandes entreprises ? Quels autres leviers de politique industrielle peuvent

être utilisés sur la base de l'expérience d'autres pays et quel serait leur impact potentiel sur les recettes fiscales et la croissance économique ?

c. Quel a été l'impact relatif des principaux leviers de politique industrielle que sont les avantages fiscaux et les subventions aux intrants sur les PME et les grandes entreprises ?

d. Évaluation des réformes relatives à la facilité de faire des affaires (« *Ease of Doing Business* ») et études comparatives : qu'est-ce qui a fonctionné et qu'est-ce qui n'a pas fonctionné ?

e. Quelle a été la création nette relative d'emplois par les grandes entreprises et les PME, l'impact des sociétés multinationales sur les entreprises nationales et leur productivité, les gains de transferts de technologie émanant des sociétés multinationales ? Quels sont les impacts relatifs des investissements directs étrangers et des investissements nationaux sur la productivité, la production et la création d'emplois ?

f. Pourquoi la chaîne de valeur mondiale (CVM) échappe-t-elle à l'Inde ? Que peut-on faire pour associer l'Inde à la CVM ? Comment les pays ont-ils établi et renforcé les liens avec la CMV et quels enseignements peut-on en tirer pour l'Inde ?

5. Finances publiques

a. Évaluation du fédéralisme fiscal de l'Inde par rapport à celui d'autres grandes démocraties.

b. Évaluation du système d'imposition direct et indirect de l'Inde sous divers aspects : descriptions et comparaisons, réaction/élasticité à différents instruments de politique fiscale et indicateurs de croissance, impact des mesures informelles sur les recettes fiscales, distorsions des incitations résultant de mesures d'application offensives, etc.

c. Évolution du régime de subventions de l'Inde : quelle est son efficacité/efficience relative dans le temps et par rapport à d'autres ?

d. Comment les différents régimes et taux de la taxe sur les produits et services se situent-ils les uns par rapport aux autres en termes de mobilisation des recettes, de bénéfices des entreprises et de croissance économique ?

6. Réformes des marchés des facteurs de production

a. Quel a été l'impact des dispositions essentielles de la loi sur les conflits du travail (« *Industrial Disputes Act* »), en particulier celles relatives aux départs, et la performance relative de l'Uttar Pradesh et d'autres États qui ont disposé de seuils plus élevés ?

b. Quels ont été les coûts liés à la multiplicité des réglementations du travail et aux exigences de conformité/dépôt/déclaration qui en découlent ? Quel a été le principal facteur de dissuasion au rassemblement de milliers d'employés dans un régime unique, en particulier dans des secteurs comme le textile ?

c. Quelles sont les principales sources de distorsion sur le marché foncier, quels ont été leurs effets respectifs, quelles sont les options politiques

possibles pour éliminer ces distorsions et comment se situent-elles les unes par rapport aux autres ?

7. Marchés informels

   a. Étude descriptive de l'économie informelle de l'Inde et de ses externalités positives et négatives. Comment a-t-elle évolué depuis la libéralisation du début des années 1990 et comment se situe-t-elle par rapport à celle des pays homologues ? Perspectives sur la manière de s'attaquer à l'économie informelle et de réduire son rôle : faut-il plutôt forcer les entreprises à aller vers l'économie formelle ou encourager les nouvelles entreprises à démarrer dans l'économie formelle ?

   b. Comment les marchés formels et informels interagissent-ils entre eux ? Quel est l'impact des grandes entreprises sur la productivité et la croissance des entreprises sur les marchés informels ?

8. Urbanisation

   a. Quels sont les coûts associés au faible coefficient d'occupation des sols (COS) de l'Inde, et à l'expansion urbaine qui en résulte en termes d'accessibilité au logement, de déplacements urbains et de pollution de l'environnement ? Quel est le coût en termes de productivité et de croissance urbaines lié au faible COS ?

   b. Évaluation des recettes de l'impôt foncier par rapport à celles des pays homologues. Quelles sont les sources possibles de revenus pour les gouvernements locaux en Inde et quel est leur potentiel respectif ? Quel est le potentiel de financement par récupération des plus-values et d'achat de COS ?

   c. Ampleur et importance du problème du caractère abordable ou non des logements en Inde et ses conséquences économiques. Quelles sont les réponses politiques découlant de l'expérience d'autres pays ? La rénovation urbaine s'est-elle accompagnée d'une gentrification et quel est son impact sur l'inclusion ? Quels sont les principaux facteurs influençant les prix de l'immobilier en zones urbaines ? Comment le programme indien de logements publics urbains a-t-il fonctionné : quel a été son impact et quelles comparaisons peut-on faire avec d'autres pays ? Quels sont les impacts relatifs des principaux instruments du type mandats, augmentation des COS, logement public ou cession de terrains publics sur les prix des logements et sur le parc de logements abordables ?

   d. Évaluation de la fourniture de services publics urbains en Inde par rapport aux pays homologues ? Quel a été le coût des externalités dues aux embouteillages, à la pollution de l'air, à l'interruption de l'approvisionnement en eau, aux installations d'égouts défectueuses et aux rejets dans les rivières, aux dépôts de déchets solides à ciel ouvert, etc. ?

   e. Comment le système indien de taxe foncière se situe-t-il par rapport à celui des pays homologues ? Comment a-t-il réagi aux diverses politiques et à l'augmentation des infrastructures ? Quelles sont les variations des taux d'imposition entre les villes indiennes, et quels enseignements peut-on en tirer en matière de bonnes pratiques et de tendances positives ?

Annexe B. Sujets de préoccupation pour les décideurs

Combien de questions ci-dessous sont susceptibles de trouver une réponse dans les RCT ? Et, au contraire, combien d'entre elles auraient pu trouver une réponse dans le cadre d'ethnographies, d'études comparatives et d'autres formes de recherches rigoureuses ?

1. Sujets de préoccupation pour les conseillers municipaux et les maires, supposant un bon rapport coût/efficacité tout en étant faisables sur les plans politique et administratif :

   a. Quelles seraient les tranches d'imposition les plus efficaces, les moins distorsives et les plus simples pour les impôts fonciers ?

   b. Quelles innovations peuvent réduire la sous-estimation par les inspecteurs des impôts – autocertification, inspections aléatoires internes, inspections aléatoires externalisées, nouvelles enquêtes périodiques, technologie ? Quels seraient les dispositifs de délégation de pouvoirs les plus efficaces pour les estimations fiscales ?

   c. Quel est le mécanisme d'incitation le plus approprié pour les inspecteurs des impôts afin de maximiser le recouvrement ?

   d. Quelles sont les approches possibles pour améliorer l'efficacité du recouvrement des impôts : récompenses, sanctions, encouragements, regroupement, mobilisation communautaire, humiliation, etc. ? Quelles sont les plus rentables ?

   e. Comment atténuer l'opposition politique (révolte, évitement fiscal) à l'introduction de nouvelles catégories de taxes : enlèvement des ordures à domicile, taxe routière sur les carburants, péages urbains, etc. ?

   f. Comment rationaliser et simplifier le processus d'agrément des constructions : où l'autocertification peut-elle être rendue suffisante, et où n'est-ce pas le cas ? Quelle innovation de processus peut-elle permettre de contrôler les infractions à la réglementation sur la construction ? Quel serait le dispositif de délégation de pouvoirs le plus efficace et le moins distorsif pour les agréments de construction ?

   g. Comment décourager la fraude sur les systèmes de raccordement d'eau, d'égouts et d'électricité et réduire le nombre de biens échappant à la facturation ? Quels sont les moyens les plus efficaces pour décourager les fraudes ?

   h. Des sommes considérables ont été investies dans les réseaux d'eau et d'égouts, mais seule une petite proportion de personnes finit par s'y raccorder en raison des coûts élevés de raccordement et d'autres obstacles. Comment faire en sorte que les ménages se raccordent aux réseaux d'eau et d'égouts une fois ces réseaux installés, sans compromettre le recouvrement des recettes ? Comment peut-on alléger les obstacles d'accès à ces réseaux ?

   i. Quels seraient les programmes de maintenance préventive les plus pratiques (pas nécessairement les meilleurs sur le plan technique) pour

les canalisations, les réseaux d'égouts, les moteurs, les routes, l'éclairage public, etc. ?

j. Comment réduire les embouteillages, en particulier sur certains tronçons routiers et à certains moments de la journée ? Ou comment réduire les embouteillages aux heures d'ouverture et de fermeture des écoles ?

k. Quelle est la meilleure stratégie possible pour l'adoption d'une planification par étapes axée sur le transport ? Par exemple, des COS plus élevés, mais autour de quelles stations de transport ou de quels lieux spécifiques ?

l. Comment développer l'utilisation des transports publics ? Comment décourager l'utilisation des véhicules privés ? Comment encourager un plus grand nombre de personnes à utiliser les transports publics ?

m. Quelle devrait être la stratégie la plus efficace pour allouer le petit nombre de logements sociaux construits chaque année entre les nombreuses demandes concurrentes : tirage au sort, établissement de critères, etc. ? Comment s'assurer que les gens ne vendent pas le logement et ne retournent pas dans des bidonvilles ?

n. Quelle serait la stratégie la plus efficace pour le déploiement des éboueurs : longueur des rues et des égouts, lieux de ramassage des ordures, etc. ?

o. Quel devrait être le calendrier de gestion des stocks le plus efficace pour les entrepôts municipaux : produits sanitaires, fournitures pour l'éclairage public, pièces de rechange pour les réseaux publics, etc. ?

p. Comment devrait être conçu le calendrier d'inspection le plus efficace (et le moins distorsif) pour les agents de recouvrement et les éboueurs employés par les municipalités ?

q. Quelle est la stratégie la plus appropriée pour récompenser les agents de recouvrement, les éboueurs, les inspecteurs des bâtiments, etc. ?

r. Quelles sont les stratégies les plus efficaces pour réduire les déchets jetés sur la voie publique ? Des poubelles ont été placées dans les lieux publics, mais les gens ne les utilisent toujours pas : alors, que faut-il faire ?

s. Comment prévenir la défécation en plein air ? Comment prévenir la miction dans les rues ? Comment s'assurer que les toilettes publiques installées sont bien utilisées ?

t. Quelle est la meilleure approche possible pour gérer les toilettes publiques ? Quelle stratégie permettrait de maximiser l'utilisation des toilettes publiques ? Comment la communauté peut-elle être mobilisée pour améliorer l'utilisation des toilettes ?

u. Quel dispositif permettrait de rationaliser la vente ambulante : réglementations, incitations, « coups de pouce » ?

v. Comment devraient être conçus les audits réalisés par des tiers pour les travaux d'ingénierie, les contrôles de présence des enseignants/fonctionnaires et l'exécution d'activités spécifiques, etc. afin qu'ils soient les plus rentables ? Quelle est la taille minimale des échantillons, la fréquence des visites et la portée des examens/inspections pour une action suffisamment dissuasive ?

w. Comment améliorer la sécurité routière – les lieux, horaires et causes des accidents sont très concentrés et spécifiques – recours à des mesures d'encouragement, des patrouilles ciblées, etc. ?

2. Questions préoccupantes pour le collecteur de district/chef du gouvernement de comté :

a. Quelles devraient être mes trois grandes priorités en matière de santé, d'éducation, d'agriculture, de moyens de subsistance et de protection sociale ? Quels sont les résultats que je devrais surveiller pour chacune de ces priorités et comment le faire ?

b. Quelles devraient être mes priorités en matière de développement rural ? Quels sont les résultats à surveiller, que faut-il faire et comment (pour chacune de ces priorités) ?

c. Quelle devrait être ma priorité en matière de développement urbain et industriel ? Quels sont les quelques résultats que je devrais surveiller pour chacun de ces domaines ? Que faut-il faire et comment (pour chacun) ?

d. Quels sont les 5 à 10 projets d'infrastructure que je devrais surveiller périodiquement ? À quelle fréquence ?

e. Comment puis-je minimiser les pertes lors des différents transferts ? Existe-t-il une solution technologique ? Existe-t-il une option de réingénierie des processus ? Est-il possible d'appliquer des mesures d'encouragement ? Les pertes sont-elles concentrées sur une série d'opérations dans la longue chaîne de transferts ?

f. Quelles sont les trois meilleures utilisations possibles d'un budget annuel de 100 millions de roupies consacré à l'innovation ?

g. Quels sont les problèmes liés aux systèmes d'achats dans les différents services ? Comment rendre les achats plus transparents et plus rentables ?

h. Quel est le moyen le plus efficace de contrôler mes services (et les programmes phares) ? Quelle devrait être la fréquence des contrôles des agents aux différents niveaux ? Quel devrait être le programme de contrôle pour chaque niveau, et comment assurer un suivi efficace ?

i. Comment puis-je utiliser au mieux les données pour suivre les activités de terrain et l'avancement de la mise en œuvre du programme ?

j. Comment optimiser l'efficacité de mes inspections sur le terrain ?

k. Comment minimiser les absences non autorisées au niveau des agents ?

l. Le collecteur de district préside de nombreux comités. Comment établir les priorités de travail entre les différents comités ?

m. Comment puis-je rendre le système de règlement des plaintes dans mon collectorat plus efficace : trouver l'équilibre en veillant à ce qu'il soit véritablement une solution de dernier recours pour les citoyens et non une solution de premier recours ?

n. Comment gérer le plus efficacement possible les audits réalisés par des tiers pour les travaux d'ingénierie, les contrôles de présence des enseignants/fonctionnaires, l'exécution d'activités, etc. ? Quelle est la taille minimale des

échantillons, la fréquence des visites et la portée des examens/inspections pour une action suffisamment dissuasive ?

o. Comment motiver mes employés (changer le discours/le langage) ? Comment tirer profit de leurs forces ? Comment les responsabiliser pour qu'ils s'approprient leur travail ?

p. Quel est le bon niveau de délégation de responsabilités aux différents chefs de service, c'est-à-dire un niveau qui favorise l'efficacité et l'appropriation sans engendrer trop de problèmes ?

q. Comment améliorer les acquis d'apprentissage ? Pour atteindre cet objectif, quelles sont les bonnes utilisations des roupies disponibles chaque année pour des dépenses discrétionnaires ? En général, quelles sont les bonnes utilisations des budgets de dépenses discrétionnaires pour les différents départements ?

r. Quelles devraient être mes initiatives phares et sur quelle base les identifier ?

s. Quelles sont les meilleures stratégies d'inspection pour les responsables des secteurs de la santé et de l'éducation à différents niveaux ? Une liste de contrôle peut-elle fonctionner ? Quelle est la liste de contrôle de second choix la plus crédible et la plus complète ?

t. Comment exploiter au mieux les visites sur le terrain des agents de vulgarisation agricole et d'autres services (qui offrent des services de conseil) ?

u. Comment dois-je rationaliser l'affectation de mon personnel : nombre adéquat de personnes qui traitent les priorités dans un secteur ?

v. Quel est le système le plus efficace pour la coordination interservices ?

w. Comment faire pour mobiliser des ressources afin que mes bureaux extérieurs locaux, dont les ressources sont limitées, disposent de suffisamment d'argent pour faire face à leurs dépenses ordinaires non salariales ?

3. Questions préoccupantes pour les décideurs politiques et le département de la santé :

a. Quelle est la meilleure utilisation de mon budget santé non pré-affecté ? Comment dois-je répartir ce budget entre les activités primaires, secondaires, tertiaires, l'enseignement médical et d'autres activités ?

b. Mes ressources en matière d'enseignants, de médecins et de personnel sont-elles affectées de manière optimale ? Quelle devrait être la meilleure politique d'affectation du personnel : critères pour les mutations, système de points ? Quelle est la meilleure stratégie d'affectation basée sur les besoins fonctionnels : médecins avec une formation de base, spécialistes ? Comment éviter les mutations/affectations *ad hoc* ?

c. Quel devrait être le système d'affectation et d'orientation le plus efficace pour les places en écoles de médecine et les places dans des fonctions paramédicales ?

d. Quelles sont les réglementations à réformer dans le domaine médical, et comment, dans un premier temps, définir celles qui sont susceptibles de générer les bénéfices les plus évidents et les plus importants, et celles qui sont réalisables ?

e. Comment réglementer le plus efficacement possible les établissements d'enseignement médical et hôpitaux privés ? Comment réglementer les pratiques médicales : quelles sont les actions qui relèvent des associations professionnelles et celles qui relèvent des gouvernements ? Comment déléguer les pouvoirs de réglementation ?

f. Comment intégrer les prestataires privés dans le système d'orientation ? Devrait-il y avoir un dispositif de régulation : engagement sur les tarifs à un niveau régional avec une certaine périodicité ?

g. Comment améliorer la qualité des conseils de traitement dans les centres de santé primaire ? Les protocoles de traitement devraient-ils être obligatoires ? Si oui, comment les faire appliquer ?

h. Quelle est la stratégie la plus efficace de saisie des données pour les dossiers médicaux électroniques ? Quels sont les domaines qui se prêtent le mieux à la digitalisation et à l'automatisation des flux de travail ? Quelle devrait être l'automatisation des flux de travail pour chaque initiative ?

i. Comment améliorer mes protocoles de lutte contre les épidémies : quelle réorganisation des processus, quel degré d'automatisation, quel niveau de délégation ?

j. Compte tenu de mes ressources, quelle devrait être ma stratégie en matière de maladies non transmissibles ? Jusqu'où les services gouvernementaux devraient-ils aller au-delà du dépistage ?

k. Comment améliorer l'efficacité des formations ? Quelle devrait être la fréquence des formations ? Quel devrait être leur contenu ? à quel niveau géographique les formations seraient-elles les plus efficaces ? Les formations devraient-elles être complétées par un mentorat ? Qui devraient être les mentors ?

l. Quelles incitations dois-je mettre en place pour les infirmières/agents sanitaires et sociaux agréés pour les accouchements en structures médicales et la vaccination : incitations en nature, points ou incitations financières ? Si elles sont financières, quel est le montant le plus susceptible d'être rentable ? Ou bien dois-je laisser les détails au libre arbitre des États (ou aux districts), et les tenir responsables des résultats liés à une mesure d'incitation globale ?

m. Comment concevoir la transition la moins distorsive et la plus pratique entre le budget par poste et le budget axé sur les résultats ? Quelle est la meilleure stratégie possible pour amorcer le passage de l'actuelle budgétisation par poste à la budgétisation axée sur les résultats ? Quels domaines ou parties du budget se prêtent le mieux au démarrage du processus ?

n. Dans quels domaines dois-je promouvoir des partenariats avec le secteur privé et quelle doit être la stratégie pour chacun d'eux ?

o. Comment intégrer dans le système les praticiens non agréés et peu qualifiés (ou « charlatans »), qui constituent, dans la grande majorité des cas, les premiers points de contact avec le secteur du soin ? Comment devrait être conçu leur module de formation ?

p. Comment rationaliser les mesures disciplinaires et simplifier les actions en justice concernant le personnel des services de santé ?

q. Quels changements peuvent être apportés à la politique de recrutement afin de la rendre plus efficace, transparente et crédible ?

r. Quels sont les changements à apporter à la politique des achats ? Comment la rendre plus transparente ? Quelle devrait être la délégation des pouvoirs la plus efficace en matière d'achats ? Comment gérer plus efficacement la chaîne d'approvisionnement en médicaments et en fournitures ? Quels achats devraient être délégués au niveau local ?

4. Les problèmes du politicien (on se sent mal rien que d'y penser…) :

a. Comment gérer un système scolaire dans lequel les enseignants ne sont là que pour enseigner, tout en faisant fonctionner une démocratie avec des recensements, des sondages, des élections, des catastrophes à répétition, etc. qui requièrent leur participation ?

b. Comment gérer une unité juridictionnelle démocratique dans laquelle personne ne veut payer d'impôts, mais où tout le monde veut des services publics de classe mondiale ?

c. Comment réglementer les contrats et les marchés avec un système de régulation en sous-effectif chronique et mal équipé ?

d. Comment réussir à faire des réformes radicales avec un État si faible qu'il ne peut pas gérer des réfectoires ?

e. Comment décentraliser et responsabiliser les fonctionnaires dans un système où la corruption et les incitations perverses sont omniprésentes ?

f. Comment supprimer les subventions inefficaces et génératrices de distorsions, et mettre en place un programme plus efficace et efficient sans provoquer de réactions politiques hostiles de tous bords ?

g. Comment savons-nous que les transferts monétaires sont un substitut plus efficace (pour la réalisation de son objectif) aux transferts en nature ?

h. Comment faire ce qui est le mieux pour le pays tout en réussissant à gagner les élections ?

Entretien avec Ila Patnaik

Ila Patnaik, vous êtes actuellement professeure au National Institute of Public Finance and Policy à New Delhi. Avant cela, vous étiez conseillère économique principale du gouvernement indien. Cela vous confère une double compétence pour évaluer le rôle des RCT dans l'élaboration des politiques en Inde. Pourquoi l'Inde est-elle devenue un lieu de prédilection pour les RCT ?

L'Inde est pratiquement devenue le centre névralgique des études RCT dans les pays en développement. Les données de l'American Economic Association concernant les RCT recensent 247 RCT réalisées en Inde depuis 2012. C'est le nombre le plus élevé dans le monde après les États-Unis. Sur ces 247 RCT, 137 ont été financées par l'Abdul Latif Jameel Poverty Action Lab (J-PAL). Les chercheurs préfèrent l'Inde aux autres pays pauvres, ce qui peut s'expliquer par le fait que l'anglais est couramment parlé et écrit dans la majeure partie du pays, et qu'il y règne un niveau de paix et de sécurité acceptable par rapport à d'autres pays en développement, où la violence et les interruptions de travail régulières peuvent venir perturber les expérimentations qui ne durent que quelques mois en raison de l'incertitude politique. En outre, une étude est moins coûteuse à réaliser en Inde que dans un pays riche. Des sommes très modestes peuvent suffire pour couvrir le coût du programme, le recrutement d'évaluateurs pour l'enquête, ainsi que les incitations financières pour les sujets afin de garantir la participation.

Les chercheurs viennent donc en Inde pour mener leurs expériences de recherche. La démarche est essentiellement conditionnée par l'offre plutôt que par la demande. En effet, les études ne sont pas induites par les problèmes auxquels les décideurs politiques sont confrontés aujourd'hui, mais par l'obligation, pour les chercheurs, de rédiger des articles pour leur doctorat ou pour des publications. La tendance est à créer de nouvelles interventions, plutôt que d'essayer d'évaluer l'impact d'une même intervention dans des régions, des communautés et des cadres de gouvernance différents.

Lorsque des chercheurs étrangers issus des meilleures universités et disposant de milliers de dollars expliquent aux bureaucrates des gouvernements locaux qu'ils peuvent investir des sommes non négligeables dans un programme qui permettra aux responsables concernés de valoriser leur image et de résoudre certains des problèmes auxquels ils font face, il est bien souvent difficile pour les bureaucrates de résister à de tels avantages. Les autorités locales « soutiennent » souvent ces études en accueillant les chercheurs, en les aidant à établir des contacts ou en fermant les yeux sur l'expérience. Les chercheurs indiens se précipitent fréquemment pour devenir administrateurs des RCT pour le compte des chercheurs étrangers. Ce sont eux qui gèrent le projet sur le terrain tandis que le chercheur international apporte les fonds et dirige le projet. Les chercheurs indiens sont parfois co-auteurs, mais ils restent généralement de simples employés.

Qu'en est-il des considérations éthiques ?

L'Inde n'a pas non plus établi de loi ou de réglementation concernant la façon dont les RCT doivent être réalisées et contrôlées. Les expériences qu'il n'est pas possible de mener aux États-Unis (par exemple, celles faisant intervenir des personnes mineures, impliquant la géolocalisation des individus ou l'absence de consentement éclairé) sont effectuées en Inde. En Inde (contrairement à de nombreux autres pays pauvres, par exemple), comme il n'existe pas d'organisme de régulation ou d'institution pour délivrer des autorisations pour la réalisation d'expérimentations, la procédure exige généralement l'approbation de l'université à laquelle le chercheur est affilié, mais personne, dans le pays, ne voit ce qui se passe. Les visiteurs, qui sont parfois des étudiants en doctorat, viennent souvent à titre personnel avec des visas touristiques, sans s'inscrire en tant que chercheur ou sans disposer d'autorisation explicite pour les travaux qu'ils mènent. Il n'existe donc aucun dispositif de contrôle des expériences réalisées en Inde ou sur des citoyens indiens. Bien souvent, on ne demande pas le consentement du sujet et cette violation des droits n'est même pas consignée. Et si elle l'est, le sujet est généralement une personne pauvre qui est prête à renoncer à sa vie privée et à céder ses informations personnelles et l'accès à sa personne ou à son entreprise contre une somme dérisoire, peu coûteuse en particulier pour les projets financés par des fonds internes.

Compte tenu des retombées pour les gouvernements, les universités et les organismes de financement, il n'est dans l'intérêt d'aucun groupe de protéger les sujets contre les expériences intrusives des chercheurs. Certains aspects sont donc très discutables en termes d'éthique, comme donner des comprimés à des enfants en bas âge, géolocaliser des étudiants ou donner des avis non scientifiques aux agriculteurs.

En 2017, le gouvernement indien a promulgué une loi sur les essais cliniques, afin d'éviter les pratiques non éthiques dans ceux qui ont lieu en Inde. Mais il reste nécessaire que le parlement indien vote également une loi et crée un organisme de régulation, ainsi qu'un cadre pour la réalisation de RCT dans le domaine des sciences sociales.

Quelle est l'utilité des RCT pour les décideurs politiques en Inde ?

Les RCT en économie du développement illustrent ce que nous pourrions appeler *jugaad* en Inde, c'est-à-dire une façon de « faire sortir quelque chose du système » sans résoudre vraiment les problèmes fondamentaux en matière de croissance économique ou de capacité de l'État. Le terme *jugaad* renvoie à une méthode permettant de résoudre un problème sans traiter les raisons qui sont à l'origine de ce problème, une méthode qui ne constitue en général pas une solution universelle. C'est un moyen qui, d'une manière ou d'une autre, offre une solution efficace à un problème, un « contournement », une « solution de fortune » qui évite de régler les questions fondamentales liées à la conception ou de se demander pourquoi le problème est apparu initialement. C'est une

solution qui doit être mise en œuvre de manière répétée parce que l'on s'attaque rarement à la racine du mal et qui fonctionne pour un problème donné, à un endroit et à un moment donnés.

Les RCT, dont on connaît la faible validité externe, ne sont pas très utiles pour prendre des décisions politiques centralisées. On sait bien que les affirmations causales d'une RCT manquent de validité externe, mais trop souvent, auteurs et lecteurs se laissent aller à une généralisation. C'est un problème particulièrement important en Inde. L'Inde est une économie à une échelle continentale, caractérisée par une diversité extrême au sein du pays. Pour bien appréhender la situation dans celui-ci, il faut comprendre que le rapport entre les régions les plus favorisées et les moins favorisées de l'Inde est comparable au rapport entre les pays les plus riches et les plus pauvres d'une zone comprenant l'Amérique latine et l'Afrique réunies. Le terme « pays » renvoyant dans l'esprit à une unité organisationnelle naturelle, nous avons tendance à faire preuve de davantage de précautions lorsque nous appliquons les résultats d'un article sur la Tanzanie, par exemple, à l'élaboration d'une politique au Chili. Mais nous sommes plus enclins à penser qu'une étude publiée en Tanzanie devrait influencer la réflexion globale à l'échelle de ce pays. De nombreux chercheurs et décideurs politiques ont ainsi tendance à extrapoler sans se poser de questions les résultats d'une RCT concernant une région de l'Inde à d'autres régions du pays. Et c'est généralement une source de problèmes. Les auteurs comme les lecteurs doivent être plus attentifs à la manière dont les articles sont écrits, afin de mieux cerner les implications d'un projet de recherche donné.

Dans chaque pays, à tout moment, certaines questions viennent occuper une place importante dans la réflexion sur les politiques. Ces dernières années en Inde, les problématiques liées à la réglementation des banques et de l'économie, à l'inflation, à la politique des taux de change et au système juridique ont par exemple occupé le devant de la scène. Dans la plupart de ces domaines, l'utilité des RCT est limitée. Elles sont utiles plutôt pour les responsables des organes de régulation bien structurés. À titre d'exemple, l'US Securities and Exchange Commission, l'organisme fédéral américain de réglementation et de contrôle des marchés financiers, mène des expérimentations lorsqu'elle réfléchit à la façon de faire évoluer la réglementation. Mais en Inde, nos problèmes sont plus basiques. Il n'existe pas d'institution de régulation responsable, intégrée dans un État de droit et disposant d'un système de pouvoirs et de contre-pouvoirs permettant de compenser les incitations individuelles au sein de l'organisation. La réflexion actuelle en matière de politiques économiques est focalisée sur des mécanismes de droit administratif qui permettraient de freiner des responsables exerçant un pouvoir coercitif. Dans ce contexte, les RCT ont un impact limité sur la façon d'appréhender les questions importantes. L'allocation disproportionnée des ressources humaines de la communauté économique en faveur des RCT a contribué à réduire la contribution des économistes au processus d'élaboration des politiques.

Quel est l'impact des RCT sur la profession d'économiste en Inde ?

L'un des moteurs importants du développement économique est la présence d'une communauté d'économistes compétents au sein du pays. Je suis un peu désabusée par l'impact de la révolution des RCT sur cette jeune communauté. Comme je l'ai indiqué plus haut, l'intérêt disproportionné porté aux problèmes qui peuvent être résolus grâce aux RCT est délétère pour une communauté de chercheurs capables de travailler sur des questions vraiment importantes. Les publications relatives aux RCT nécessitent des financements de grande ampleur selon les standards indiens. Pour l'essentiel, ces fonds proviennent de sources étrangères, ce qui a créé une forme de dépendance particulière : les chercheurs, à qui l'on a dit que seuls des travaux concernant les RCT pouvaient être publiés, sont contraints de solliciter les réseaux internationaux qui sont à même de réunir les vastes ressources requises pour mener ces recherches. Cela a nui à l'authenticité des réflexions, et entravé la capacité à observer le monde et à se poser les questions importantes. Au lieu d'avoir une seule distorsion (le décalage entre les intérêts des éditeurs, des évaluateurs scientifiques et des chercheurs), il existe maintenant trois distorsions supplémentaires (seules les problématiques auxquelles on peut répondre à l'aide des RCT sont éligibles aux recherches, seules les questions qui intéressent les bailleurs de fonds étrangers sont prises en considération, et seuls les sujets pour lesquels il existe de tels réseaux internationaux peuvent être traités). Les publications sur les RCT se concentrent majoritairement sur des questions très limitées (par exemple : l'assiduité des enseignants s'améliore-t-elle lorsque l'enseignant doit poster un selfie chaque matin), et cela s'est fait au détriment d'une intellectualisation plus large de la jeune communauté des chercheurs en économie du développement. La capacité intellectuelle à penser l'économie et à réfléchir aux problématiques de l'Inde a décliné en échange des perspectives de carrière offertes par une publication dans des revues scientifiques de renom.

La réalisation de RCT constitue-t-elle une allocation efficace de ressources limitées au service de l'économie indienne ?

Si j'étais un planificateur central allouant des ressources pour mieux servir l'économie en Inde, il me semble évident que la meilleure façon d'optimiser le potentiel serait de créer davantage de données de toutes natures, et des données de meilleure qualité. En Inde, nous ne disposons notamment pas de données de qualité sur le produit intérieur brut (PIB), la consommation ou l'emploi. Il y a même un débat autour du nombre réel des personnes sans emploi dans le pays. Élaborer des politiques en l'absence de données revient à conduire avec un bandeau sur les yeux. Du point de vue d'un économiste, il serait plus utile d'utiliser les ressources pour mesurer l'économie, l'emploi, le marché du travail, etc. Nous ne disposons d'aucune mesure relative au secteur informel, pas plus qu'à l'innovation, aux compétences ou à la productivité, alors que ce sont des données essentielles pour élaborer des politiques et stimuler la croissance. Nous naviguons à vue lorsqu'il s'agit de l'économie dans son ensemble. Si,

grâce aux RCT, nous connaissons le comportement des agents sur les marchés locaux du crédit, nous avons beaucoup de mal à mesurer l'évolution des taux d'intérêt sur les marchés informels du crédit.

Les données économiques concernant les ménages en Inde sont limitées. À titre d'exemple, les premières données de panel sur les ménages, qui reposent sur trois observations annuelles d'un échantillon de ménages, ne datent que de ces dernières années. La production d'un tel ensemble de données nécessite des ressources comparables à celles de quelques RCT, et va permettre de créer un corpus complet de nouvelles connaissances sur l'Inde. Une grande partie de ces connaissances seront descriptives, et d'autres quasi expérimentales. Cet investissement semble constituer une meilleure utilisation des ressources, en comparaison de quelques RCT qui coûteraient la même somme pour peu de résultats et une recherche non reproductible. Les données de panel permettent une réplication et font naître une concurrence entre les chercheurs (pour un investissement modeste), ce que les RCT ne sont pas capables de faire dans une telle mesure. Les données de panel offrent une mesure durable de faits simples et élémentaires concernant le pays (par exemple sur la participation des femmes à l'emploi), ce que les RCT ne permettent pas. Les données de panel sont beaucoup plus précieuses pour le développement d'une communauté de recherche indienne pertinente pour les politiques. Je comprends bien en quoi les problèmes de représentation des organisations gouvernementales et philanthropiques combinés à la recherche de solutions très à la mode ces dernières années ont incité à « découvrir ce qui fonctionne », au détriment d'une construction plus complexe d'un savoir en sciences humaines et sociales indispensable à l'élaboration des politiques. Mais, après ces longues années de « révolution des RCT », je plaide en faveur d'une réorientation vers des approches plus traditionnelles.

Bibliographie

ACEMOGLU D., ROBINSON J., 2006 – *Economic origins of dictatorship and democracy*. Cambridge, Cambridge University Press.

ACEMOGLU D., 2010 – Theory, General Equilibrium, and Political Economy in Development Economics. *Journal of Economic Perspectives*, 24 (3) : 17-32.

ADAMS V. (ed.), 2016 – *Metrics: What Counts in Global Health*. Durham/ Londres, Duke University Press.

AFD/EUDN (éd.), 2012 – Unease in Evaluation: What Are the Lessons to be Drawn from the Development Experience? *Revue d'économie du développement*, 20 (4), numéro spécial.

AHMED H. M., MITCHELL M., HEDT B., 2010 – National Implementation of Integrated Management of Childhood Illness (IMCI): Policy Constraints and Strategies. *Health Policy*, 96 (2) : 128-133.

AIKEN A. M., DAVEY C., HARGREAVES J. R., HAYES R. J., 2015 – Re-analysis of Health and Educational Impacts of a School-based Deworming Programme in Western Kenya: A Pure Replication. *International Journal of Epidemiology*, 44 (5) : 1572-1580.

AKRICH M., STRUM S., CALLON M., LATOUR B., 2013 – *Sociologie de la traduction : textes fondateurs*. Paris, Presses des Mines.

ALCOTT H., 2015 – Site Selection Bias in Program Evaluation. *The Quarterly Journal of Economics*, 130 (3) : 1117-1165.

ALFONSI L., BANDIERA O., BASSI V., BURGESS R., RASUL I., SULAIMAN M., VITALI A., 2017 – *Tackling Youth Unemployment: Evidence from a Labor Market Experiment in Uganda*. Working Paper, Private Enterprise Development in Low Income Countries (PEDL) Programme, Londres, DFID.

ALIK LAGRANGE A., RAVALLION M., 2019 – Estimating Within-Group Spillover Effects Using a Cluster Randomization: Knowledge Diffusion in Rural India. *Journal of Applied Econometrics*, 34 : 110-128.

ALKEMA L., CHOU D., HOGAN D., ZHANG S., MOLLER A., GEMMILL A., FAT D. M., BOERMA T., 2016 – Global, Regional, and National Levels and Trends in Maternal Mortality between 1990 and 2015, with Scenario-based Projections to 2030: A

Systematic Analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet*, 387 (10017) : 462-474.

ALKIN M. C., 2004 – *Evaluation Roots: Tracing Theorists' Views and Influences*. Beverly Hills, Sage Publications.

ANDERSON D. M., CHARLES K. K., REES D. I., 2018 – *Public Health Efforts and the Decline in Urban Mortality.* NBER Working Paper, 25027.

ANDRÉS L., BRICEÑO B., CHASE C., ECHENIQUE J. A., 2017 – Sanitation and Externalities: Evidence from Early Childhood Health in Rural India. *Journal of Water, Sanitation and Hygiene for Development*, 7 (2) : 272-289.

ANDREWS M., PRITCHETT L., WOOLCOCK M., 2012 – *Escaping Capability Traps through Problem-Driven Iterative Adaptation*. Center for Global Development Working Paper n° 299.

ANDREWS M., PRITCHETT L., WOOLCOCK M., 2017 – *Building State Capability: Evidence, Analysis, Action*. Oxford, Oxford University Press.

ANGELL M., 1997 – Editorial: The Ethics of Clinical Research in the Third World. *New England Journal of Medicine*, 337 (12) : 847-849.

ANGELUCCI M., KARLAN D., ZINMAN J., 2015 – Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7 (1) : 151-182.

ANGRIST J., KRUEGER A., 1999 – « Empirical Strategies in Labor Economics ». *In* ASHENFELTER O., CARD D. (eds) : *Handbook of Labor Economics*, vol. 3. Amsterdam, North-Holland : 1277-1366.

ANGRIST J. D., PISCHKE J.-S., 2010 – The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24 (2) : 3-30.

ANGRIST J., IMBENS G., RUBIN D., 1996 – Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, XCI : 444-455.

ANGULO SALAZAR L., 2013 – « The Social Costs of Microfinance and Over-indebtedness for Women ». *In* GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds) : *Microfinance, Debt and Over-indebtedness. Juggling with Money*. Londres, Routledge : 232-252.

ANSAH E. K., NARH-BANA S., ASIAMAH S., DZORDZORDZI V., BIANTEY K., DICKSON K., GYAPONG J. O., KORAM K. A., GREENWOOD B. M., MILLS A., WHITTY C. J. M., 2009 – Effect of Removing Direct Payment for Health Care on Utilisation and Health Outcomes in Ghanaian Children: A Randomised Controlled Trial. *PLoS Medicine*, 6 (1) : 48-58.

ARDILLY P., TILLÉ Y., 2006 – *Sampling Methods: Exercises and Solutions*. Basingstoke, Springer.

ARISTOTE, s.d. – *La rhétorique.* http://remacle.org/bloodwolf/philosophes/Aristote/rheto1.htm

ARMENDÀRIZ B., MORDUCH J., 2010 – *The Economics of Microfinance*. Cambridge, MIT Press.

ARNOLD B. F., HOGAN D. R., COLFORD J. M., HUBBARD A. E., 2011 – Simulation Methods to Estimate Design Power: An Overview for Applied Research. *BMC Medical Research Methodology*, 11 (1) : 94.

ARNOLD B. F., NULL C., LUBY S. P., COLFORD J. M. Jr., 2018 – Implications of WASH Benefits Trials for Water and Sanitation. Authors' Reply. *The Lancet Global Health*, 6 (6) : e616-e617.

ARNOTT R., STIGLITZ J. E., 1988 – Randomization with Asymmetric Information. *RAND Journal of Economics*, 19 (3) : 344-362.

ARSHAD A., SALAM R. A., LASSI Z. S., DAS J. K., NAQVI I., BHUTTA Z. A., 2014 – Community Based Interventions for the Prevention and Control of Tuberculosis. *Infectious Diseases of Poverty*, 3 (1) : 1-10. https://doi.org/10.1186/2049-9957-3-27

ARUNACHALAM R. S., 2011 – *The Journey of Indian Micro-finance: Lessons for the Future*. Chennai, Aapti Publications.

ASHENFELTER O. C., 1983 – Determining Participation in Income-Tested Social Programs. *Journal of the American Statistical Association*, 78 (383) : 517-525.

ASHENFELTER O. C., CARD D., 1985 – Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *Review of Economics and Statistics*, 67 (4) : 648-660.

ATHEY S., IMBENS G. W., 2017 – « The Econometrics of Randomized Experiments ». *In* : BANERJEE A., DUFLO E. (eds) : *The Handbook of Economic Field Experiments*. Amsterdam, North-Holland : 73-140.

ATHEY S., IMBENS G. W., 2018 – *Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption*. NBER Working Paper, 24963.

ATHEY S., IMBENS G. W., 2019 – Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11 (1) : 685-725.

ATHEY S., CHETTY R., IMBENS G., HYUNSEUNG K., 2016 – *Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index*, inédit. https://arxiv.org/abs/1603.09326

ATTANASIO O., AUGSBURG B., DE HAAS R., Fitzsimons E., HARMGART H., 2015 – The Impacts of Microfinance: Evidence from Joint-liability Lending in Mongolia. *American Economic Journal: Applied Economics*, 7 (1) : 90-122.

ATUN R. A., BENNETT S., DURAN A., 2008 – When Do Vertical (Stand-alone) Programmes Have a Place in Health Systems? *WHO European Ministerial Conference on Health Systems*, 1-28.

AUGSBURG B., DE HAAS R., HARMGART H., MEGHIR C., 2015 – The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7 (1) : 183-203.

BAELE S., 2013 – The Ethics of New Development Economics: Is the Experimental Approach to Development Economics Morally Wrong? *Journal of Philosophical Economics*, 7 (1) : 1-42.

BAHADUR R. R., SAVAGE L. J., 1956 – The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *Annals of Mathematical Statistics*, 27 (4) : 1115-1122.

BAIRD S., BOHREN A., MCINTOSH C., ÖZLER B., 2017 – Optimal Design of Experiments in the Presence of Interference. *Review of Economics and Statistics*, 100 (5) : 844-860.

BALASUNDARAM F. M. T., MURALIDHARAN A., RAMAN V., MOSLER V., MOSLER H. J., 2019 – *Promoting Latrine Use in Karnataka, India Using the RANAS Approach to Behaviour Change*. New Delhi, International Initiative for Impact Evaluation, 3rd Grantee Final.

BALDASSARRI D., ABASCAL M. 2017 – Field Experiments across the Social Sciences. *Annual Review of Sociology*, 43 : 41-73.

BAMBERGER M., RAO V., WOOLCOCK M., 2010 – *Using Mixed Methods in Monitoring and Evaluation, Experiences from International Development*. World Bank Policy Research Working Paper, 5245.

BANDIERA O., BURGESS R., DAS N., GULESCI S., RASUL I., SULAIMAN M., 2017 – Labor Markets and Poverty in Village Economies. *Quarterly Journal of Economics*, 132 (2) : 811-870.

BANERJEE A., 2006 – Making Aid Work. How to Fight Global Poverty. Effectively, *Boston Review*.

BANERJEE A. (ed.), 2007 – *Making Aid Work*. Cambridge/Londres, MIT Press.

BANERJEE A., 2013 – *The J-PAL Story: A Decade of Partnerships.* https://www.youtube.com/watch?v=AkC9tVUptM4&list=PL5Dr5MK6NSso3iEqn6BDu8OzyMFyLwiNE&index=5

BANERJEE A., DUFLO E., 2009 – The Experimental Approach to Development Economics. *Annual Review of Economics*, 1 : 151-178.

BANERJEE A., DUFLO E., 2011 – *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York, Public Affairs.

BANERJEE A., DUFLO E., 2014 – « The Experimental Approach to Development Economics ». *In* TEELE D. L. (ed.) : *Field Experiments and Their Critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven/Londres, Yale University Press : 78-114.

BANERJEE A., DUFLO E., 2017 – Pushing Evidence-Based Policymaking for the Poor. *Livemint*, 16 octobre. https://www.livemint.com/Opinion/nYjG4JP2ve6YpeXkMb3AHJ/Pushing-evidencebased-policymaking-for-the-poor.html

BANERJEE A. V., DUFLO E., 2019 – *Good Economics for Hard Times: Better Answers to our Biggest Problems.* Londres, Penguin Books UK.

BANERJEE A., HE R., 2008 – « Making Aid Work ». *In* EASTERLY W. (ed.) : *Reinventing Foreign Aid*. Cambridge, The MIT Press : 47-92.

BANERJEE A., MULLAINATHAN S., 2010 – *The Shape of Temptation: Implications for the Economic Lives of the Poor*. NBER Working Paper, 15973.

BANERJEE A., DUFLO E., GLENNERSTER R., 2008 – Putting a Band-aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System. *Journal of the European Economic Association*, 6 : 487-500.

BANERJEE A., BANERJI R., DUFLO E., GLENNERSTER R., KHEMANI S., 2010 – Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy*, 2 : 1-30.

BANERJEE A., CHATTOPADHYAY R., DUFLO E., KENISTON D., SINGH N., 2012 – *Can Institutions Be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police*. NBER Working Paper, 17912.

BANERJEE A., DUFLO E., GOLDBERG N., KARLAN D., OSEI R., PARIENTÉ W., SHAPIRO J., THUYSBAERT B., UDRY C., 2015a – A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries. *Science*, 348 (6236) : 1-16.

BANERJEE A., DUFLO E., GLENNERSTER R., KINNAN C., 2015b – The Miracle of Microfinance? Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*, 7 (1) : 22-53.

BANERJEE A., KARLAN D., ZINMAN J., 2015c – Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal: Applied Economics*, 7 (1) : 1-21.

BANERJEE A., DUFLO E., KREMER M., 2016 – The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. Communication, *The State of Economics, The State of the World* : 42. https://economics.mit.edu/files/16473

BANERJEE A., CHASSANG S., SNOWBERG E., 2017a – « Decision Theoretic Approaches to Experiment Design and External Validity ». *In* BANERJEE A., DUFLO E. (eds) : *Handbook of Economic Field Experiments*. Vol. 1, Amsterdam, Elsevier : 141-174.

BANERJEE A., CHASSANG S., MONERO S., SNOWBERG E., 2017b – *A Theory of Experimenters*. NBER Working Paper, 23867.

BANERJEE A., DUFLO E., KENISTON D., SINGH N., 2017c – One Question for the Road, *The Indian Express*.

BANERJEE A., BREZA E., DUFLO E., KINNAN C., 2019a – *Can Microfinance Unlock a Poverty Trap for Some Entrepreneurs?* NBER Working Paper, 26346.

BANERJEE A., DUFLO E., M. KREMER, 2019b – « The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy ». *In* BASU K., ROSENBLATT D., SEPULVEDA C. P. (eds) : *State of Economics, State of the World*. Cambridge, MIT Press : 439-487.

BARDHAN P., 1984 – *Land, Labor, and Rural Poverty: Essays in Development Economics*. New York, Columbia University Press.

BARRETT C., CARTER M., 2010 – The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections. *Applied Economic Perspectives and Policy*, 32 (4) : 515-548.

BARRETT C., CARTER M., 2014 – « A Retreat from Radical Skepticism: Rebalancing Theory, Observational Data, and Randomization in Development Economics ». *In* TEELE D. L. (ed.) : *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven/Londres, Yale University Press : 58-77.

BARRETT C., CARTER M., 2020 – Finding Our Balance? Revisiting the Randomization Revolution in Development Economics Ten Years Further On. *World Development*, 127 : 104789. https://doi.org/10.1016/j.worlddev.2019.104789

BASTIAENSEN J., MARCHETTI P., 2011 – « Rural Microfinance and Agricultural Value Chains: Strategies and Perspectives of the Fondo de Desarrollo Local in Nicaragu ». *In* ARMANDARIZ B., LABIE M. (eds) : *The Handbook of Microfinance*. Londres/Singapour, World Scientific Publishing : 461-495.

BASU K., 2013 – *Shared Prosperity and the Mitigation of Poverty in Practice and in Precept*. World Bank Working Paper, 6700.

BASU K., 2014 – Randomization, Causality and the Role of Reasoned Intuition. *Oxford Development Studies*, 42 (4) : 455-472.

BATEMAN M., 2010 – *Why Doesn't Microfinance Work? The Destructive Rise of Local Neoliberalism*. Londres, Zed Books.

BATES M. A., GLENNERSTER R., 2017 – The Generalizability Puzzle. *Stanford Social Innovation Review*, 51-4. https://doi.org/10.48558/eyy5-3s89

BAUCHET J., MORDUCH J., 2019 – Paying in Pieces: A Natural Experiment on Demand for Life Insurance under Different Payment Schemes. *Journal of Development Economics*, 139 (C) : 69-77.

BAUCHET J., MORDUCH J., RAVI S., 2015 – Failure versus Displacement: Why an Innovative Anti-poverty Program Showed No Net Impact in South India. *Journal of Development Economics*, 116 (C) : 1-16.

BATES M. A., GLENNERSTER R., GUMEDE K., DUFLO E., 2012 – The Price Is Wrong. *The Journal of Field Action, Field Actions Science Reports*, numéro spécial (4). https://factsreports.revues.org/1554

BEACCO J.-C., MOIRAND S., 1995 – Autour des discours de transmission des connaissances. *Langages*, 117 : 32-53.

BEAMAN L., BENYISHAY A., MAGRUDER J., MOBARAK A. M., 2018a – *Can Network-Theory Based Targeting Increase Technology Adoption?* NBER Working Paper, 24912.

BEAMAN L., KARLAN D., THUYSBAERT B., UDRY C., 2018b – *Selection into Credit Markets: Evidence from Agriculture in Mali*. Working Paper.

Bédécarrats F., 2012 – L'impact de la microfinance : un enjeu politique au prisme de ses controverses scientifiques. *Mondes en développement*, 2 : 127-142.

Bédécarrats F., Guérin I., Roubaud F., 2013 – L'étalon-or des évaluations randomisées : du discours de la méthode à l'économie politique. *Sociologies pratiques*, 2 : 107-122.

Bédécarrats F., Guérin I., Morvant-Roux S., Roubaud F., 2019a – Estimating Microcredit Impact with Low Take-up, Contamination and Inconsistent Data. A Replication Study of Crépon, Devoto, Duflo, and Pariente. (American Economic Journal: Applied Economics, 2015). *International Journal for Re-Views in Empirical Economics*, 3. https://www.jcr-econ.org/estimating-microcredit-impact-replication/

Bédécarrats F., Guérin I., Roubaud F., 2019b – All that Glitters Is Not Gold. The Political Economy of Randomized Evaluations in Development. *Development and Change*, 50 (3) : 735-762.

Bédécarrats F., Guérin I., Morvant-Roux S., Roubaud F., 2019c – *Verifying the Internal Validity of a Flagship RCT: A Review of Crépon, Devoto, Duflo and Pariente. Rebutting the Rebuttal*. DIAL Working Paper, 2019-07B.

Bédécarrats F., Guérin I., Morvant-Roux S., Roubaud F., 2021 – Behind the Scenes of Science in Action: Tinkering with a Randomized Control Trial in Morocco. *Third World Quaterly*, 42 (11). https://doi.org/10.1080/01436597.2021.1977114

Beisel U., 2015 – Markets and Mutations: Mosquito Nets and the Politics of Disentanglement in Global Health. *Geoforum*, 66 : 146-155.

Belissa T., Bulte E., Cecchi F., Gangopadhyay S., Lensink R., 2019 – Liquidity Constraints, Informal Institutions, and the Adoption of Weather Insurance: A Randomized Controlled Trial in Ethiopia. *Journal of Development Economics*, 140 : 269-278.

Ben David D., Papell D. H., 1998 – Slowdowns and Meltdowns: Postwar Growth Evidence from 74 Countries. *The Review of Economics and Statistics*, 80 : 561-571.

Berg A., Ostry J. D., Zettelmeyer J., 2012 – What Makes Growth Sustained? *Journal of Development Economics*, 98 : 149-166.

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) : 314-327.

Berndt C., 2015 – Behavioural Economics, Experimentalism and the Marketization of Development. *Economy and Society*, 44 (4) : 567-591.

Bernhardt A., Field E., Pande R., Rigol N., 2017 – *Household Matters: Revisiting the Returns to Capital among Female Micro-entrepreneurs*. NBER Working Paper, 23358.

Berriet-Solliec M., Labarthe P., Laurent C., 2014 – Goals of Evaluation and Types of Evidence. *Evaluation*, 20 (2) : 195-213.

BERTRAND M., DJANKOV S., HANNA R., MULLAINATHAN S., 2007 – Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption. *Quarterly Journal of Economics*, 122 (4) : 1639-1676.

BERTRAND M., KARLAN D., MULLAINATHAN S., SHAFIR E., ZINMAN J., 2010 – What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment. *The Quarterly Journal of Economics*, 125 (1) : 263-306.

Beta-Blocker Heart Attack Trialists (BHAT), 1982 – A Randomized Trial of Propranolol in Patients with Acute Myocardial Infarction. *Journal of the American Medical Association*, 247 (12) : 1707-1714.

BETHLEHEM J., 2009 – *Applied Survey Methods: A Statistical Perspective.* New York, Wiley.

BHATT S., WEISS D., CAMERON E., BISANZIO D., MAPPIN B., DALRYMPLE U., BATTLE K., MOYES C. L., HENRY A., ECKHOFF P. A., WENGER E. A., BRIËT O., PENNY M. A., SMITH T. A., BENNETT A., YUKICH J., EISELE T. P., GRIFFIN J. T., FERGUS C. A., LYNCH M., LINDGREN F., COHEN J. M., MURRAY C. L. J., SMITH D. L., HAY S. I., CIBULSKIS R. E., GETHING P. W., 2015 – The Effect of Malaria Control on Plasmodium Falciparum in Africa between 2000 and 2015. *Nature*, 526 (7572) : 207-211.

BHATTACHARYYA O., REEVES S., ZWARENSTEIN M., 2009 – What Is Implementation Research? *Research on Social Work Practice*, 19 (5) : 491-502.

BHUTTA Z. A., DAS J. K., RIZVI A., GAFFEY M. F., WALKER N., HORTON S., WEBB P., LARTEY A., BLACK R. E., 2013 – Evidence-based Interventions for Improvement of Maternal and Child Nutrition: What Can Be Done and at What Cost? *The Lancet*, 382 (9890) : 452-477.

BIEHL J., PETRYNA A., BIEHL J., PETRYNA A., 2014 – Peopling Global Health, *Saúde e Sociedade*, 23 (2) : 376-389.

BISBEE J., DEHEJIA R., POP-ELECHES C., SAMII C., 2017 – Local Instruments, Global Extrapolation: External Validity of the Labor Supply. Fertility Local Average Treatment Effect. *Journal of Labor Economics*, 35 (S1) : S99-S147.

BLATTMAN C., 2008 – *Impact Evaluation 2.0.* https://www.chrisblattman.com/documents/policy/2008.ImpactEvaluation2.DFID_talk.pdf

BLATTMAN C., DERCON S., 2018 – The Impacts of Industrial and Entrepreneurial Work on Income and Health: Experimental Evidence from Ethiopia. *American Economic Journal: Applied Economics*, 10 (3) : 1-38.

BLUSTEIN J., 2005 – Toward a More Public Discussion of the Ethics of Federal Social Program Evaluation. *Journal of Policy Analysis and Management*, 24 (4) : 824-852.

BOLD T., KIMENYI M., MWABU G., NG'ANG'A 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. *Sexually Transmitted Diseases*, 40 (2) : 111-112.

Bold T., Kimenyi M., Mwabu G., Ng'ang'a A., Sandefur J., 2018 – Experimental Evidence on Scaling Up Education Reforms in Kenya. *Journal of Public Economics*, 168 : 1-20.

Bonds M. H., Rich M. L., 2018 – Integrated Health System Strengthening Can Generate Rapid Population Impacts that Can Be Replicated: Lessons from Rwanda to Madagascar. *BMJ Global Health*, 3 (5) : e000976.

Boone P., Eble A., Elbourne D., 2013 – *Risk and Evidence of Bias in Randomized Controlled Trials in Economics*. Londres, Centre for Economic Performance, LSE.

Borgerson K., 2009 – Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine. *Perspectives in Biology and Medicine*, 52 (2) : 218-233.

Bornmann L., Mutz R., 2014 – Growth Rates of Modern Science: A Bibliometric Analysis based on the Number of Publications and Cited References. *Journal of the Association of Information Science and Technology*, 66 : 2215-2222.

Bothwell L., Podolsky S. H., 2016 – The Emergence of the Randomized Controlled Trial. *The New England Journal of Medicine*, 375 (6) : 501-504.

Bothwell L., Greene J., Podolsky S., Jones D., 2016 – Assessing the Gold Standard: Lessons from the History of RCTs. *New England Journal of Medicine*, 374 (22) : 2175-2181.

Botros S., 1990 – « Equipoise, Consent and the Ethics of Randomised Clinical Trials ». *In* Byrne P. (ed.) : *Ethics and Law in Health Care and Research*. Chichester, John Wiley & Sons : 9-24.

Bouguen A., Huang Y., Kremer M., Miguel E., 2019 – Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics. *Annual Review of Economics*, 11 : 523-561.

Bouquet E., Wampfler B., Ralison E., Roesch M., 2007 – Trajectoires de crédit et vulnérabilité des ménages ruraux : le cas des Cecam de Madagascar. *Autrepart*, 4 : 157-172.

Bourdieu P., 1975 – The Specificity of the Scientific Field and the Social Conditions of the Progress of Reason. *Sociology of Science*, 14 (6) : 19-47.

Bradford-Hill A., 1965 – The Environment and Disease Association or Causation. *Proceedings of the Royal Society of Medicine*, 58 : 295-300.

Breton P., 1999 – La « préférence manipulatoire » du président du Front national. *Mots*, 58 : 101-125.

Breuer J. B., McDermott J., 2013 – Economic Depression in the World. *Journal of Macroeconomics*, 38 : 227-242.

Broadbent A., Vandenbroucke J. P., Pearce N., 2017 – Formalism or Pluralism? A Reply to Commentaries on "Causality and Causal Inference in Epidemiology". *International Journal of Epidemiology*, 45 (6) : 1841-1851.

Brodeur A., Lé M., Sangnier M., Zylberberg Y., 2016 – Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8 (1) : 1-32.

Brodeur A., Cook N., Heyes A., 2018 – *Methods Matter: P-Hacking and Causal Inference in Economics*. IZA Working Paper, 11796.

Brody C., De Hoop T., Vojtkova M., Warnock R., Dunbar M., Murthy P., Dworkin S. L., 2015 – Economic Self-help Group Programs for Improving Women's Empowerment: A Systematic Review. *Campbell Systematic Reviews*, 11 (1) : 1-182.

Brown C., Ravallion M., D. van de Walle, 2018 – A Poor Means Test? Econometric Targeting in Africa. *Journal of Development Economics*, 134 : 109-124.

Bruhn M., McKenzie D., 2009 – In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, 1 (4) : 200-232.

Brummitt C. D., Huremovic K., Pin P., Bonds M. H., Vega-Redondo F., 2017 – Contagious Disruptions and Complexity Traps in Economic Development. *Nature Human Behaviour*, 1 (9) : 665-672.

Bryce J., Requejo J. H., Moulton L. H., Ram M., Black R. E., 2013 – A Common Evaluation Framework for the African Health Initiative. *BMC Health Services Research*, 13 (2) : S10.

Buera F. J., Kaboski J. P., Shin Y., 2015 – Entrepreneurship and Financial Frictions: A Macrodevelopment Perspective. *Economics*, 7 (1) : 409-436.

Bursztyn L., Cantoni D., Yang D., Yuchtman N., Zhang Y. J., 2019 – *Persistent Political Engagement: Social Interactions and the Dynamics of Protest Movements*. NBER Conference Paper, F126621.

Burtles G., 1995 – The Case for Randomized Field Trials in Economic and Policy Research. *Journal of Economic Perspectives*, 9 (2) : 63-84.

Bylander M., 2014 – Borrowing across Borders: Migration and Microcredit in Rural Cambodia. *Development and Change*, 45 (2) : 284-307.

Cai J., Szeidl A., 2018 – Interfirm Relationships and Business Performance. *The Quarterly Journal of Economics*, 133 (3) : 1229-1282.

Cain G. G., 1975 – « Regression and Selection Models to Improve Nonexperimental Comparisons ». *In* Bennett C., Lumsdaine A. A. (eds) : *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*. New York, Academic Press : 297-317.

Cain G. G., Watts H. W., 1973 – « Summary and Overview ». *In* Cain G. G., Watts H. W. (eds) : *Income Maintenance and Labor Supply: Econometric Studies*. Chicago, Markham : 163-181.

Cain G. G., Wissoker D. A., 1990 – A Reanalyis of Marital Stability in the Seattle-Denver Income-Maintenance Experiment. *American Journal of Sociology*, 95 (5) : 1235-1269.

CALLON M., 2006a – « Pour une sociologie des controverses technologiques ». *In* AKRICH M., LATOUR B. (eds) : *Sociologie de la traduction. Textes fondateurs*. Paris, Presses des Mines : 135-157.

CALLON M., 2006b – « Quatre modèles pour décrire la dynamique de la science ». *In* AKRICH M., LATOUR B. (eds) : *Sociologie de la traduction. Textes fondateurs.* Paris, Presses des Mines : 201-251.

CAMERER C. F., DREBER A. FORSELL E., HO T. H., HUBER J., JOHANNESSON M., KIRCHLER M., ALMENBERG J., ALTMEJD A., CHAN T., 2016 – Evaluating Replicability of Laboratory Experiments in Economics. *Science*, 351 (6280) : 1433-1436.

CAMERON D. B., MISHRA A., BROWN A. N., 2016 – The Growth of Impact Evaluation for International Development: How Much Have We Learned? *Journal of Development Effectiveness*, 8 (1) : 1-21.

CAMFIELD L., DUVENDACK M., 2014 – Impact Evaluation: Are We "Off the Gold Standard"? *The European Journal of Development Research*, 26 (1) : 1-11.

CAMFIELD L., DUVENDACK M., PALMER-JONES R., 2014 – Things You Wanted to Know about Bias in Evaluations but Never Dared to Think. *IDS Bulletin*, 45 (6) : 49-64.

CAMPBELL D. T., 1974 – *Qualitative Knowing in Action Research*. Prix Kurt Lewin, Society for the Psychological Study of Social Issues, presenté à la réunion de l'American Psychological Association, Nouvelle-Orléans, 30 septembre.

CAMPBELL D. T., STANLEY J. C., 1963 – *Experimental and Quasi-Experimental Designs for Research*. New York, Houghton Mifflin Co., 2.

CAMPBELL M., FITZPATRICK R., HAINES A., KINMONTH A. L., SANDERCOCK P., SPIEGELHALTER D., TYRER P., 2000 – Framework for Design and Evaluation of Complex Interventions to Improve Health Framework for Trials of Complex Interventions. *British Medical Journal*, 321 (7262) : 694-696.

CAMPBELL N. C., MURRAY E., DARBYSHIRE J., EMERY J., FARMER A., GRIFFITHS F., GUTHRIE B., LESTER H., WILSON P., KINMONTH A. L., 2007 – Designing and Evaluating Complex Interventions to Improve Health Care. *British Medical Journal*, 334 (7591) : 455-459.

CAMPOS F., FRESE M., GOLDSTEIN M., IACOVONE L., JOHNSON H., MCKENZIE D., MENSMANN M., 2017 – Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa. *Science*, 357 (6357) : 1287-1290.

CAPLAN A. L., 2001 – « Twenty Years After: The Legacy of the Tuskegee Syphilis Study ». *In* TEAYS W., PURDY L. M. (eds) : *Bioethics, Justice and Health Care*. Belmont, Wadsworth-Thomson Learning : 231-235.

CARD D., DELLAVIGNA S., 2013 – *Nine Facts about Top Journals in Economics*. NBER Working Paper, 18665.

CARLSON R. V., BOYD K. M., WEBB D. J., 2004 – The Revision of the Declaration of Helsinki: Past, Present and Future. *British Journal of Clinical Pharmacology*, 57 (6) : 695-713.

CARTWRIGHT N., 2007 – Are RCTs the Gold Standard? *BioSocieties*, 2 (1) : 11-20.

CARTWRIGHT N., 2010 – What Are Randomised Controlled Trials Good For? *Philosophical Studies*, 147 (1) : 59.

CARTWRIGHT N., HARDIE J., 2012 – *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford, Oxford University Press.

CARTWRIGHT N., MUNRO E., 2010 – The Limitations of Randomized Controlled Trials in Predicting Effectiveness. *Journal of Evaluation in Clinical Practice*, 16 (2) : 260-266.

CARUSO B. A., SCLAR G. D., ROUTRAY P., NAGEL C., MAJORIN F., SOLA S., KOEHNE W., DESHAY R., UDAIPURIA S., WILLIAMS R., CLASEN T., 2019 – *Impacts of a multi-level intervention, Sundara Grama, on latrine use and safe disposal of child faeces in rural Odisha, India*. New Delhi, International Initiative for Impact Evaluation, 3rd Grantee Final Report.

CASABURI L., WILLIS J., 2018 – Time versus State in Insurance: Experimental Evidence from Contract Farming in Kenya. *American Economic Review*, 108 (12) : 3778-3813.

CASE A., DEATON A., 2015 – Rising Mortality and Morbidity among Midlife White Non-Hispanics in 21st century America. *Proceedings of the National Academy of Sciences of the USA*, 112 (49) : 15078-15083.

CASE A., PAXSON C., 2008 – Stature and Status: Height, Ability, and Labor Market Outcomes. *Journal of Political Economy*, 116 (3) : 499-532.

CASEY K., GLENNERSTER R., MIGUEL E., VOORS M., 2018 – *Skills versus Voice in Local Development*. Non publié.

CEDERLÖF G., 1997 – *Bonds Lost: Subordination, Conflict and Mobilisation in Rural South India c. 1900–1970*. New Delhi, Manohar.

Centre for Global Development, 2006 – *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Rapport du groupe de travail sur les lacunes en matière d'évaluation, Washington, Center for Global Development.

CHABÉ-FERRET S., 2018 – *An Approach Combining Theory, Simulations and Empirics Provides Evidence of Regularities in the Bias of Observational Methods*. Toulouse, Toulouse School of Economics.

CHAKRAVARTY E. F., FRIES J. F., 2006 – Science as Experiment; Science as Observation. *Nature Clinical Practice Rheumatology*, 2 (6) : 286.

CHASSANG S., MIQUEL P. I., SNOWBERG E., 2012 – Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments, *American Economic Review*, 102 (4) : 1279-1309.

CHATTERJEE P., 2008 – Clinical Trials in India: Ethical Concerns. *Bulletin of the World Health Organization*, 86 (8) : 581-582.

CHAUHAN K., SCHMIDT W. P., AUNGER R., GOPALAN B., SAXENA D., YASHOBANT S., PATWARDHAN V., CURTIS V., 2019 – *The 5 Star Toilet Campaign: Improving*

Toilet Use in Gujarat. New Delhi, International Initiative for Impact Evaluation, 3rd Grantee Final Report.

Chavance B., Labrousse A., 2018 – Institutions and Science: The Contest about Pluralism in Economics in France. *Review of Political Economy*, 30 (2) : 190-209.

Chayanov A. V., 1966 [1925] – *The Theory of Peasant Economy*. Homewood, Richard Irwin for the American Economic Association.

Chen S., Mu R., Ravallion M., 2009 – Are There Lasting Impacts of Aid to Poor Areas? Evidence from Rural China. *Journal of Public Economics*, 93 : 512-528.

Chenery H., Ahluwalia M., Bell C., Duloy J., Jolly R., 1979 – *Redistribution with Growth*. New York, Oxford University Press.

Chernozhukov V., Demirer M., Duflo E., Fernandez-Val I., 2018 – *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments*. Working Paper, Cambridge, National Bureau of Economic Research.

Cherrier B., 2019 – Weekly Lecture Was on 'What Should Come First: Theory or Data?' So Here's Tweetstorm on the History of Quantitative Economics. Twitter, 13 mars. twitter.com/Undercoverhist/status/1105851715461570560

Chetty R., Hendren N., Jones M. R., Porter S. R., 2019 – *Race and Economic Opportunity in the United States: An Intergenerational Perspective*. NBER Working Paper, 24441.

Childress J. F., Faden R. R., Gaare R. D., Gostin L. O., Kahn J., Bonnie R. J., Nieburg P., 2002 – Public Health Ethics: Mapping the Terrain. *Journal of Law, Medicine & Ethics*, 30 (2) : 170-178.

Chong A., La Porta R., Lopez-de-Silanes F., Shleifer A., 2014 – Letter Grading Government Efficiency. *Journal of the European Economics Association*, 12 (2) : 277-298.

Chouliaraki L., Fairclough N., 2010 – Critical Discourse Analysis in Organizational Studies: Towards an Integrationist Methodology, *Journal of Management Studies*, 47 (6) : 1213-1218.

Christensen G., Miguel E., 2018 – Transparency, Reproducibility, the Credibility of Economics Research. *Journal of Economic Literature*, 56 (3) : 920-980.

Christiano L., Eichenbaum M., Rebelo S., 2011 – When Is the Government Spending Multiplier Large? *Journal of Political Economy*, 119 : 78-121.

Clasen T., Boisson S., Routray P., Torondel B., Bell M., Cumming O., Ensink J., Freeman M., Jenkins M., Odagiri M., 2014 – Effectiveness of a Rural Sanitation Programme on Diarrhoea, Soil-Transmitted Helminth Infection, and Child Malnutrition in Odisha, India: A Cluster-Randomised Trial. *The Lancet Global Health*, 2 (11) : e645-e653.

Cling J.-P., Razafindrakoto M., Roubaud F., 2003 – *New International Poverty Reduction Strategies*. Londres/New York, Routledge.

CLING J.-P., LAGRÉE S., RAZAFINDRAKOTO M., ROUBAUD F., 2014 – *The Informal Economy in Developing Countries*. Londres/New York, Routledge.

COFFEY D., SPEARS D., 2017 – *Where India Goes: Abandoned Toilets, Stunted Development and the Costs of Caste*. Londres, Harper Collins.

COFFEY D., SPEARS D., 2018 – Implications of WASH Benefits Trials for Water and Sanitation. *The Lancet Global Health*, 6 (6) : 615.

COFFEY D., DEATON A., DRÈZE J., SPEARS D., TAROZZI A., 2013 – Stunting among Children: Facts and Implications. *Economic and Political Weekly*, 48 (34) : 68-69.

COHEN J., DUPAS P., 2010 – Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment. *Quarterly Journal of Economics*, 125 (1) : 1-45.

COHEN J., EASTERLY W., 2010 – *What Works in Development? Thinking Big and Thinking Small*. Washington, Brookings Institution Press.

COLLINS D., MORDUCH J., RUTHERFORD S., RUTHVEN O., 2009 – *Portfolios of the Poor: How the World's Poor Live on $2 a Day*. Princeton, Princeton University Press.

CONCATO J., 2012 – Is It Time for Medicine-Based Evidence? *The Journal of the American Medical Association*, 307 (15) : 1641-1643.

CONCATO J., 2013 – Study Design and 'Evidence' in Patient-Oriented Research. *American Journal of Respiratory and Critical Care Medicine*, 187 (11) : 1167-1172.

CONCATO J., HORWITZ R. I., 2004 – Beyond Randomised versus Observational Studies, *The Lancet*, 363 (9422) : 1660-1661.

CONCATO J., HORWITZ R. I., 2018 – Randomized Trials and Evidence in Medicine: A Commentary on Deaton and Cartwright. *Social Science & Medicine*, 210 : 32-36.

CONCATO J., SHAH N., HORWITZ R., 2000 – Randomized Controlled Trials, Observational Studies, and the Hierarchy of Research Design. *New England Journal of Medicine*, 342 (25) : 1887-1892.

Congress of the United States, SCOFSOPA [Senate, Committee on Finance, Subcommittee on Public Assistance], 1978 – *Welfare Research and Experimentation: Hearings Before the Subcommittee on Public Assistance of the Committee on Finance, United States Senate, Ninety-Fifth Congress, Second Session, November 15, 16, and 17*. Washington, Congress of the United States.

CONLISK J., 1973 – Choice of Response Functional Form in Designing Subsidy Experiments. *Econometrica*, 41 (4) : 643-656.

CONLISK J., WATTS H., 1969 – A Model for Optimizing Experimental Designs for Estimating Response Surfaces. *American Statistical Association Proceedings, Social Statistics Section* : 150-156.

Cook T. D., 2018 – Twenty-six Assumptions that Have to Be Met If Single Random Assignment Experiments Are to Warrant 'Gold Standard' Status: A Commentary on Deaton and Cartwright. *Social Science and Medicine*, 210 : 37-40.

Cook T. D., Campbell D. T., 1979 – *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago, Rand McNally College Publishing Company.

Copestake J., Bhalotra S., Johnson S., 2001 – Assessing the Impact of Microcredit: A Zambian Case Study. *The Journal of Development Studies*, 37 (4) : 81-100.

Copestake J., Dawson P., Fanning J.-P., McKay A., Wright-Revolledo K., 2005 – Monitoring the Diversity of the Poverty Outreach and Impact of Microfinance: A Comparison of Methods Using Data from Peru. *Development Policy Review*, 23 (6) : 703-723.

Copestake J., Johnson S., Cabello M., Goodwin-Groen R., Gravesteijn R., Humberstone J., Nino-Zarazua M., Titus M., 2016 – Towards a Plural History of Microfinance. *Canadian Journal of Development Studies/Revue canadienne d'études du développement*, 37 (3) : 279-297.

Cornia G., Jolly R., Stewart F. (eds.), 1987 – *Adjustment with a Human Face: Protecting the Vulnerable and Promoting Growth*. Oxford, Oxford University Press.

Council for International Organizations and Medical Sciences (CIOMS), 2002 – *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Genève, World Health Organization.

Coville A., Vivalt E., 2017 – How Often Should We Believe Positive Results? Assessing the Credibility of Research Findings in Development Economics, non publié.

Cox D. R., 1958 – *Planning of Experiments*. New York, Wiley.

Cox D. R., Reid N., 2000 – *The Theory of the Design of Experiments*. New York, Chapman and Hall.

Crépon B., Devoto F., Duflo E., Parienté W., 2015 – Estimating the Impact of Microcredit on Those Who Take It up: Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics*, 7 (1) : 123-150.

Crépon B., Devoto F., Duflo E., Parienté W., 2019 – *"Verifying the Internal Validity of a Flagship RCT: A Review of Crépon, Devoto, Duflo and Parienté": A Rejoinder*. DIAL Working Paper, 07A.

Croke K., Hicks J. H., Hsu E., Kremer M., Miguel E., 2016 – *Does Mass Deworming Affect Child Nutrition? Meta-analysis, Cost-Effectiveness, and Statistical Power*. NBER Working Paper, 22382.

Cronbach L., 1982 – *Designing Evaluations of Educational and Social Programs*. San Francisco, Jossey Bass.

Crucifix C., Morvant-Roux S., 2018 – « Fragmented Rural Communities: The Faenas of Prospera at the Interface of Community Cooperation and State

Dependency ». *In* Balen M. E., Fotta M. (eds) : *Money from the Government in Latin America: Conditional Cash Transfer Programs and Rural Lives*. Londres/ New York, Routledge : 123-148.

Cull R., Morduch J., 2018 – « Microfinance and Economic Development ». *In* Beck T., Levine R. (eds) : *Handbook of Finance and Development*. Cheltenham, Edward Elgar : 550-572.

Cull R., Ehrbeck T., Holle N., 2014 – *Financial Inclusion and Development: Recent Impact Evidence*. Focus note 92, Washington, CGAP.

Cull R., Demirgüç-Kunt A., Morduch J., 2018 – The Microfinance Business Model: Enduring Subsidy and Modest Profit. *The World Bank Economic Review*, 32 (2) : 221-244.

Cumming O., Curtis V., 2018 – Implications of WASH Benefits Trials for Water and Sanitation. *The Lancet Global Health*, 6 (6) : e613-e614.

Cumming O., Arnold B. F., Ban R., Clasen T., Esteves Mills J., Freeman M. C., Gordon B., Guiteras R., Howard G., Hunter P. R., 2019 – The Implications of Three Major New Trials for the Effect of Water, Sanitation and Hygiene on Childhood Diarrhea and Stunting – A Consensus Statement. *BMC Medicine*, 17. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1410-x

Cutler D., Miller G., 2005 – The Role of Public Health Improvements in Health Advances: The Twentieth-century United States. *Demography*, 42 (1) : 1-22.

Czibor E., Jimenez-Gomez D., List J. A., 2019 – *The Dozen Things Experimental Economists Should Do (More of)*. NBER Working Paper, 25451.

Dahal M., Fiala N., 2020 – What Do We Know about the Impact of Microfinance? The Problems of Power and Precision. *World Development*, 128 : 104773.

DARPA SCORE, s.d. – *Systematizing Confidence in Open Research and Evidence (SCORE)*, Arlington, DARPA. https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence

Dasgupta P., Marglin S., Sen A., 1972 – *Guidelines for Project Evaluation*. Vienne, United Nations Industrial Development Organization.

Datta L.-E., 1994 – Paradigm Wars: A Basis for Peaceful Co-existence and Beyond. *New Directions for Program Evaluation*, 61 : 53-70.

Davey C., Aiken A. M., Hayes R. J., Hargreaves J. R., 2015 – Re-analysis of Health and Educational Impacts of a School-Based Deworming Programme in Western Kenya: A Statistical Replication of a Cluster Quasi-Randomized Stepped-Wedge Trial. *International Journal of Epidemiology*, 44 (5) : 1581-1592.

Davis D., Holt C. A., 1993 – *Experimental Economics*. Princeton, Princeton University Press.

Days S. J., Altman D. G., 2000 – Blinding in Clinical Trials and Other Studies. *British Medical Journal*, 321 : 504.

De Dickert N. W., Emanuel E. J., 2015 – « Ethics in Cardiovascular Medicine ». *In* Mann D. L., Zipes D. P., Libby P., Bonow R. O (eds) : *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*. Philadelphie, Elsevier Saunders : 29-34.

de Souza Leão L., Eyal G., 2020 – Searching under the Streetlight: A Historical Perspective on the Rise of Randomistas. *World Development*, 127 : 104781.

Deaton A., 1997 – *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington, World Bank.

Deaton A., 2010a – Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48 (2) : 424-455.

Deaton A., 2010b – Understanding the Mechanisms of Economic Development. *Journal of Economic Perspectives*, 24 (3) : 3-16.

Deaton A., 2012 – *Searching for Answers with Randomized Experiments*. New York, Development Research Institute. https://www.youtube.com/watch?v=yiqbmiEalRU

Deaton A., 2013a – The Financial Crisis and the Wellbeing of Americans. *Oxford Economic Papers*, 64 (1) : 1-26.

Deaton A., 2013b – *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, Princeton University Press.

Deaton A., 2015 – The Logic of Effective Altruism. *Boston Review.* http://bostonreview.net/forum/logic-effective-altruism/angus-deaton-response-effective-altruism

Deaton A., Cartwright N., 2018 – Understanding and Misunderstanding Randomized Controlled Trials. *Social Science and Medicine*, 210 : 2-21.

Deaton A., Dupriez O., 2011 – Purchasing Power Parity Exchange Rates for the Global Poor. *American Economic Journal: Applied Economics*, 3 (2) : 137-166.

Deaton A., Stone A. A., 2016 – Understanding Context Effects for a Measure of Life Evaluation: How Responses Matter. *Oxford Economic Papers*, 68 (4) : 861-870.

Dehejia R., Morduch J., Montgomery H., 2012 – Do Interest Rates Matter? Credit Demand in the Dhaka Slums. *Journal of Development Economics*, 47 (2) : 437-499.

Dehejia R., Pop-Eleches C., Samii C., 2019 – *From Local to Global: External Validity in a Fertility Natural Experiment*. NBER Working Paper, 21459.

DellaVigna S., Pope D., 2018a – Predicting Experimental Results: Who Knows What? *Journal of Political Economy*, 126 (6) : 2410-2456.

DellaVigna S., Pope D., 2018b – What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies*, 85 (2) : 1029-1069.

DellaVigna S., Pope D., Vivalt E., 2019 – Predict Science to Improve Science. *Science*, 366 (6464) : 428-429.

DEMIRGUC-KUNT A., KLAPPER L., SINGER D., 2017 – *Financial Inclusion and Inclusive Growth: A Review of Recent Empirical Evidence*. Washington, World Bank.

DESROSIÈRES A., 1998 – *The Politics Of Large Numbers: A History Of Statistical Reasoning*. Cambridge, Harvard University Press.

DESROSIÈRES A., 2013a – *Pour une sociologie historique de la quantification : l'argument statistique*. Paris, Presses des Mines. https://books.openedition.org/pressesmines/901

DESROSIÈRES A., 2013b – *Gouverner par les nombres : l'argument statistique II*. Paris, Presses des Mines. https://books.openedition.org/pressesmines/341

Development Assistance Committee, 2010 – *Glossary of Key Terms in Evaluation and Results Based Management*. Paris, OECD Editions.

Department for International Development (DFID), 2012 – *Broadening the Range of Designs and Methods for Impact Evaluations*. DFID Working Paper, 38.

DHALIWAL I., HANNA R., 2013 – *Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India*. NBER Working Paper, 20482.

DHALIWAL I., OLKEN B., 2018 – *Announcing J-PAL's Policy Insights*, 1. https://www.povertyactionlab.org/blog/5-10-18/announcing-j-pals-policy-insights

DI TILLIO A., OTTAVIANI M., SØRENSEN P. N., 2017 – Persuasion Bias in Science: Can Economics Help? *Economic Journal*, 127 (605) : F266-F304.

DILLON A., KARLAN D., UDRY C., ZINMAN J., 2020 – Good Identification, Meet Good Data. *World Development*, 127 : 104796.

DIMOVA R., 2019 – A Debate that Fatigues…: To Randomise or Not to Randomise ; What's the Real Question? *The European Journal of Development Research*, 31 (2) : 163-168.

DOKOVA M., 2016 – The Role of Captatio Benevolentiae in the Interaction between the Speaker and His Audience in Antiquity and Today. *Systasis*, 29. https://www.academia.edu/30224939/The_role_of_captatio_benevolentiae_in_the_interaction_between_the_speaker_and_his_audience_in_Antiquity_and_today

DOLIGEZ F., 2002 – Microfinance et dynamiques économiques : quels effets après dix ans d'innovations financières ? *Revue tiers monde*, 43 (172), 783-808.

DONOVAN K., 2018 – The Rise of the Randomistas: On the Experimental Turn in International Aid. *Economy and Society*, 47 (1) : 27-58.

DOOLITTLE F. C., TRAEGER L., 1990 – *Implementing the National Jtpa Study*. New York, Manpower Demonstration Research Corporation.

DRÈZE J., 2018a – Evidence, Policy, and Politics. *Ideas for India*. https://www.ideasforindia.in/topics/miscellany/evidence-policy-and-politics.html

DRÈZE J., 2018b – Evidence, Policy and Politics: A Commentary on Deaton and Cartwright. *Social Science & Medicine*, 210 : 45-47.

DUBNER S. J., 2018 – *Is the Protestant Work Ethic Real?* Freakonomics Podcast, épisode 360. http://freakonomics.com/podcast/religiosity/

DUFLO E., 2009 – *Expérience, science et lutte contre la pauvreté*. Paris, Fayard.

DUFLO E., 2017 – Richard T. Ely Lecture: The Economist as Plumber. *American Economic Review: Papers and Proceedings*, 107 (5) : 1-26.

DUFLO E., KREMER M., 2003 – *Use of Randomization in the Evaluation of Development Effectiveness*. Washington, World Bank Operations Evaluation Department Conference on Evaluation and Development Effectiveness, 15-16 juillet. https://economics.mit.edu/files/765

DUFLO E., GLENNERSTER R., KREMER M., 2004 – Randomized Evaluations of Interventions in Social Service Delivery. *Development Outreach*, 6 (1) : 26-29.

DUFLO E., GLENNERSTER R., KREMER M., 2011 – « Using Randomization in Development Economics Research: A Toolkit ». *In* SCHULTZ T., STRAUSS J. (eds) : *Handbook of Development Economics*. Vol. 4, Amsterdam, North-Holland : 3895-3962.

DUFLO E., HANNA R., RYAN S. P., 2012 – Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102 : 1241-1278.

DUFLO E., GREENSTONE M., PANDE R., RYAN N., 2013 – Truth-Telling by Third-Party Auditors and the Response of Polluting Firms: Experimental Evidence from India. *The Quarterly Journal of Economics*, 128 (4) : 1-49.

DUFLO E., DUPAS P., KREMER M., 2015a – School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools. *Journal of Public Economics*, 123 : 92-110.

DUFLO E., GREENSTONE M., GUITERAS R., CLASEN T., 2015b – *Toilets Can Work: Short and Medium Run Health Impacts of Addressing Complementarities and Externalities in Water and Sanitation*. NBER Working Paper, 21521.

DUMEZ H., JEUNEMAÎTRE A., 2005 – La démarche narrative en économie. *Revue économique*, 56 (4) : 983-1006.

DUNCAN G., HUSTON A., WEISNER T., 2007 – *Higher Ground: New Hope for the Working Poor and Their Children*. New York, Russell Sage.

DUPAS P., KARLAN D., ROBINSON J., UBFAL D., 2018 – Banking the Unbanked? Evidence from Three Countries. *American Economic Journal: Applied Economics*, 10 (2) : 257-297.

DURBIN J., 1954 – Errors in Variables. *Review of the International Statistical Institute*, 22 : 23-32.

DUTEIL-MOUGEL C., 2005 – Les mécanismes persuasifs des textes politiques. *Corpus*, 4. http://corpus.revues.org/357

DUVENDACK M., PALMER-JONES R., 2012 – High Noon for Microfinance Impact Evaluations: Re-investigating the Evidence from Bangladesh. *The Journal of Development Studies*, 48 (12) : 1864-1880.

DUVENDACK M., PALMER-JONES R., COPESTAKE J. G., Hooper L., LOKE Y., RAO N., 2011 – *What is the Evidence of the Impact of Microfinance on the*

Well-being of Poor People? Londres, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

EASTERLY W., 2001 – *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics*. Cambridge, The MIT Press.

EASTERLY W., 2007 – *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. Oxford, Oxford University Press.

EASTERLY W., 2012 – If Christopher Columbus Had Been Funded by Gates. *NYU Development Research Institute Blog*. https://nyudri.wordpress.com/2012/10/15/if-christopher-columbus-had-been-funded-by-gates/

EASTERLY W., 2013 – *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. New York, Basic Books.

EASTERLY W., 2019 – *In Search of Reforms for Growth: New Stylized Facts on Policy and Growth Outcomes*. NBER Working Paper, 26318.

EGIL F., 2015 – Les objectifs de développement durable, nouveau « palais de cristal » ? *Politique africaine*, 4 : 99-120.

EL-SADR W. M., PHILIP N. M., JUSTMAN J. E., 2014 – Letting HIV Transform Academia. Embracing Implementation Science. *New England Journal of Medicine*, 370 (18) : 1679-1681.

ELYACHAR J., 2006 – *Markets of Dispossession: NGOs, Economic Development, and the State in Cairo*. Durham, Duke University Press.

ELYACHAR J., 2012 – Next Practices: Knowledge, Infrastructure, and Public Goods at the Bottom of the Pyramid. *Public Culture*, 24.1 (66) : 109-129.

ENCISCO A. L., 2019 – Acaba el clientelar Prospera ; surge el programa Becas Benito Juárez, *La Jornada*, 30 janvier : 32. https://www.jornada.com.mx/2019/01/30/sociedad/032n1soc

EVANS D., 2016 – That Zero Effect May Not Mean What You Think It Means, and Other Lessons from Recent Educational Research. *Development Impact Blog*, Washington, World Bank. https://blogs.worldbank.org/impactevaluations/zero-effect-may-not-mean-what-you-think-it-means-and-other-lessons-recent-educational-research

EVANS D., POPOVA A., 2016 – *What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews*. World Bank Policy Research Working Paper, 7203.

FARMER P., MURRAY M., HEDT-GAUTHIER B., 2013 – Clinical Trials in Global Health Equity. *Lancet Global Health blog*. http://globalhealth.thelancet.com/2013/07/08/clinical-trials-and-global-health-equity

FASSIN D., 2010 – *La raison humanitaire. Une histoire morale du temps présent*. Paris, Gallimard/Seuil.

FAULKNER W. N., 2014 – A Critical Analysis of a Randomized Controlled Trial Evaluation in Mexico: Norm, Mistake or Exemplar? *Evaluation*, 20 (2) : 230-243.

FAVEREAU J., 2016 – On the Analogy between Field Experiments in Economics and Clinical Trials in Medicine. *Journal of Economic Methodology*, 23 (2) : 203-222.

FEINSTEIN A. R., HORWITZ R. I., 1997 – Problems in the "Evidence" of "Evidence-Based Medicine". *The American Journal of Medicine*, 103 (6) : 529-535.

FERGUSON J., 1990 – *The Anti-Politics Machine: Development, Depoliticization and Bureaucratic Power in Lesotho*. Cambridge, Cambridge University Press.

FERGUSON J., 2015 – *Give a Man a Fish: Reflections on the New Politics of Distribution*. Durham/Duke, Duke University Press.

FEW S., 2009 – Statistical Narrative. Telling Compelling Stories with Numbers. *Visual Business Intelligence Newsletter*, juillet-août : 1-10.

FIELD E., PANDE R. PAPP J., RIGOL N., 2013 – Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India. *American Economic Review*, 103 (6) : 2196-2226.

FIENNES C., 2018 – Funders Start Assessing Their Own Performance. To Understand What a Charity is Achieving, You Must Understand What Good Research Looks Like. *Financial Times*, 27 novembre.

FILMER D., PRITCHETT L., 1999 – What Education Production Functions Really Show: A Positive Theory of Education Expenditures. *Economics of Education Review*, 18 : 223-239.

FINE B., JOHNSTON D., SANTOS A. C., VAN WAEYENBERGE E., 2016 – Nudging or Fudging: The World Development Report 2015. *Development and Change*, 47 (4) : 640-663.

FINKELSTEIN A., TAUBMAN S., 2015 – Randomize Evaluations to Improve Health Care Delivery. *Science*, 347 (6223) : 720-722.

FISHER R. A, 1926 – The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33 : 503-513.

FISHER R. A., 1935 – *The Design of Experiments*. Londres, Oliver and Boyd.

FISHER R. A., 1960 – *The Design of Experiments*. Seventh Edition, Edinburgh, Oliver and Boyd.

FISZBEIN A., SCHADY N., 2010 – *Conditional Cash Transfers for Attacking Present and Future Poverty*. Washington, World Bank.

Food and Drug Administration, 2010 – *Adaptive Design Clinical Trials for Drugs and Biologics*. Washington, FDA.

FORD D. G., 2002 – Teaching Anecdotally. *College Research*, 50 (3) : 114-115.

FOURCADE M., OLLION E., ALGAN Y., 2015 – The Superiority of Economists. *Journal of Economic Perspectives*, 29 (1) : 89-114.

FRAKER T., MAYNARD R., 1987 – The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs. *Journal of Human Resources*, 22 (2) : 194-227.

FRANCO A., MALHOTRA N., SIMONOVITS G., 2014 – Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science*, 345 (6203) : 1502-1505.

FREEDMAN B., 1987 – Equipoise and the Ethics of Clinical Research. *The New England Journal of Medicine*, 317 (3) : 141-145.

FREEDMAN D. A., 1991 – Statistical Models and Shoe Leather. *Sociological Methodology*, 21 : 291-313.

FREEDMAN D. H., 2010 – Lies, Damned Lies and Medical Science. *The Atlantic*, 306 (4) : 76-84. https://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/

FRENCH J., BLAIR-STEVENS C., MCVEY D., MERRITT R., 2010 – *Social Marketing and Public Health: Theory and Practice*. Oxford, Oxford University Press.

FRIEDEN T. R., 2017 – Evidence for Health Decision Making. Beyond Randomized Controlled Trials. *New England Journal of Medicine*, 377 (5) : 465-475.

FRIEDMAN M., 2009 – *Capitalism and Freedom.* Chicago, University of Chicago Press.

FRIEDMAN J., GOKUL B., 2014 – Quantifying the Hawthorne Effect. *Development Impact Blog*, Washington, World Bank.

FRIES J. F., KRISHNAN E., 2004 – Equipoise, Design Bias, and Randomized Controlled Trials: The Elusive Ethics of New Drug Development. *Arthritis Research & Therapy*, 6 (3) : R250.

GALASSO E., RAVALLION M., 2005 – Decentralized Targeting of an Anti-Poverty Program. *Journal of Public Economics*, 89 (4) : 705-727.

GALASSO E., RAVALLION M., SALVIA A., 2004 – Assisting the Transition from Workfare to Work: Argentina's *Proempleo* Experiment. *Industrial and Labor Relations Review*, 57 (5) : 128-142.

GARCHITORENA A., MILLER A. C., CORDIER L. F., RABEZA V. R., RANDRIAMANAMBINTSOA M., RAZANADRAKATO H.-T. R., HALL L., GIKIC D., HARUNA J., MCCARTY M., RANDRIANAMBININA A., THOMSON D. R., ATWOOD S., RICH M. L., MURRAY M. B., RATSIRARSON J., OUENZAR M. A., BONDS M. H., 2018 – Early Changes in Intervention Coverage and Mortality Rates Following the Implementation of an Integrated Health System Intervention in Madagascar. *BMJ Global Health*, 3 (3) : e000762.

GASS J., PRITCHETT L., 2017 – *Returns on Scholarship (versus Organizational Learning) in Development Using (mostly) Education as an Example*. Présentation à l'Université de Washington.

GAUTAM M., 2000 – *Agricultural Extension: The Kenya Experience*. OED Precis, 198, Washington, World Bank.

GEERTZ C., 1973 – *The Interpretation of Cultures.* New York, Basic Books.

GELBACH J. B., PRITCHETT L., 2002 – Is More for the Poor Less for the Poor? The Politics of Means-Tested Targeting. *The B.E. Journal of Economic Analysis & Policy*, 2 (1) : 1-28.

GELMAN A., 2018 – Benefits and Limitations of Randomized Controlled Trials: A Commentary on Deaton and Cartwright. *Social Science & Medicine*, 210 : 48-49.

GELMAN A., CARLIN J., 2014 – Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9 (6) : 641-651.

GELMAN A., TUERLINCKX F., 2000 – Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures. *Computational Statistics*, 15 : 373-390.

GERTLER P., 2004 – Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. *The American Economic Review*, 94 (2) : 336-341.

GERTLER P. J., MARTINEZ S., PREMAND P., RAWLINGS L. B., VERMEERSCH C. M. J., 2016 – *Impact Evaluation in Practice*. Washington, Inter-American Development Bank/World Bank.

GERUSO M., SPEARS D., 2018 – Neighborhood Sanitation and Infant Mortality. *American Economic Journal: Applied Economics*, 10 (2) : 125-162.

GHOSH A., GUPTA A., SPEARS D., 2014 – Are Children in West Bengal Shorter Than Children in Bangladesh? *Economic & Political Weekly*, 48 (8) : 21-24.

GIBSON J., 2019 – Are You Estimating the Right Thing? An Editor Reflects. *Applied Economic Perspectives and Policy*, 41 (3) : 329-350.

GIEDION U., ALFONSO E. A., DÍAZ Y., 2013 – *The Impact of Universal Coverage Schemes in the Developing World: A Review of the Existing Evidence*. UNICO Studies Series, 25, Washington, World Bank.

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

GLENNERSTER R., 2016 – *Not So Small. Running Randomized Evaluations.* http://runningres.com/blog/2016/5/27/not-so-small

GLENNERSTER R., POWERS S., 2016 – « Balancing Risk and Benefit. Ethical Tradeoffs in Running Randomized Evaluations ». *In* DEMARTINO G., MCCLOSKEY D. (eds) : *Oxford Handbook on Professional Economic Ethics*, Oxford, Oxford University Press : 367-401.

GLENNERSTER R., TAKAVARASHA K., 2013 – *Running Randomized Evaluations: A Practical Guide*. Princeton, Princeton University Press.

GLYNN A., KASHIN K., 2018 – Front-door Versus Back-door Adjustment with Unmeasured Confounding: Bias Formulas for Front-door and Hybrid Adjustments with Application to a Job Training Program. *Journal of the American Statistical Association*, 113 (523) : 1040-1049.

GOLDBERG J., 2014 – *The R-Word Is Not Dirty*. Washington, Center for Global Development.

GOLDBERGER A. S., MANSKI C. F., 1995 – Review Article: The Bell Curve by Herrnstein and Murray. *Journal of Economic Literature*, 33 (2) : 762-776.

GOSSET W., 1937 – Comparison between Balanced and Random Arrangements of Field Plots. *Biometrika*, 29 : 363-379.

GRASSO P. G., WASTY S. S., WEAVING R. V. (eds.), 2003 – *World Bank Operations Evaluation Department: The First 30 Years.* Washington, World Bank.

GREEN W., 1991 – *Econometric Analysis*. New York, Macmillan.

GROSH M., GLEWWE P. (eds.), 2000 – *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 years of the Living Standards Measurement Study*. Washington, World Bank.

GROSSMAN J., MACKENZIE F., 2005 – The Randomized Controlled Trial: Gold Standard, or Merely Standard? *Perspectives in Biology and Medicine*, 48 (4) : 516-534.

Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI), 1987 – Long-term Effects of Intravenous Thrombolysis in Acute Myocardial Infarction: Final Report of the GISSI Study. *Lancet*, 335 (8687) : 427-431.

GUBERT F., ROUBAUD F., 2011 – *The Impact of Microfinance Loans on Small Informal Enterprises in Madagascar: A Panel Data Analysis*. Washington, World Bank.

GUÉRIN I., KUMAR S., 2017 – Market, Freedom and the Illusions of Microcredit. Patronage, Caste, Class and Patriarchy in Rural South India. *The Journal of Development Studies*, 53 (5) : 741-754.

GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds.), 2013a – *Microfinance, Debt and Over-indebtedness: Juggling with Money*. Londres/New York, Routledge.

GUÉRIN I., ROESCH M., VENKATASUBRAMANIAN G., KUMAR S., 2013b – « The Social Meaning of Over-indebtedness and Creditworthiness in the Context of Poor Rural South Indian Households (Tamil Nadu) ». *In* GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds) : *Microfinance, Debt and Over-indebtedness: Juggling with Money*, Londres/New York, Routledge : 125-149.

GUÉRIN I., LABIE M., SERVET J.-M. (eds), 2015 – *The Crises of Microcredit*. Londres, Zed Book.

GUÉRIN I., VENKATASUBRAMANIAN G., KUMAR S., 2019 – Rethinking Saving: Indian Ceremonial Gifts as Relational and Reproductive Saving. *Journal of Cultural Economy.* https://doi.org/10.1080/17530350.2019.1583594

GUERON J., 2017 – « The Politics and Practice of Social Experiments: Seeds of a Revolution ». *In* BANERJEE A., DUFLO E. (eds) : *The Handbook of Economic Field Experiments*. Vol. 1, Amsterdam, North-Holland : 27-69.

GUERON J., ROLSTON H., 2013 – *Fighting for Reliable Evidence*. New York, Russell Sage Foundation.

GUGERTY M. K., KARLAN D., 2018 – Ten reasons not to measure impact – and what to do instead. *Stanford Social Innovation Review*, Summer issue : 41-47.

GUGERTY M. K., KREMER M., 2008 – Outside Funding and the Dynamics of Participation in Community Associations. *American Journal of Political Science*, 52 (3) : 585-602.

GULHATI C. M., 2004 – Needed: Closer Scrutiny of Clinical Trials. *Indian Journal of Medical Ethics*, 1 : 4-5.

GUYER J. I., 1997 – « Endowments and Assets: The Anthropology of Wealth and the Economics of Intrahousehold Allocation ». *In* HADDAD LAWRENCE J., HODDINOTT J., ALDERMAN H. (eds) : *Intrahousehold Resource Allocation in Developing Countries*. Baltimore, The Johns Hopkins University Press : 112-129.

HAAVELMO T., 1944 – The Probability Approach in Econometrics. *Econometrica*, 12 (Supplement) : iii-vi et 1-115.

HAHN J., TODD P., VAN DER KLAAUW W., 2001 – Identification and Estimation of Treatment Effects with a Regression-Discontinutiy Design. *Econometrica*, 69 (1) : 201-209.

HALL P., 2017 – « The Role of Interests, Institutions and Ideas in the Political Economy of Industrialized Nations ». *In* LICHBACH M., ZUCKERMAN A. (ed.) : *Comparative Politics: Rationality, Culture and Structure*. Cambridge, Cambridge University Press : 174-207.

HALPERN S. D., KARLAWISH J. H., BERLIN J. A., 2002 – The Continuing Unethical Conduct of Underpowered Clinical Trials. *Journal of the American Medical Association*, 288 (3) : 358-362.

HAM J. C., LALONDE R. J., 1990 – « Using Social Experiments to Estimate the Effect of Training on Transition Rates ». *In* HARTOG J., RIDDER G., THEEUWES J. (eds) : *Panel Data and Labor Market Studies*. Oxford, North-Holland : 157-172.

HAMMER J., 2014 – The Chief Minister Posed Questions We Couldn't Answer. *Building State Capacity Blog*, Harvard University. https://buildingstatecapability.com/2014/04/08/the-chief-minister-posed-questions-we-couldnt-answer/

HAMMER J., 2017 – *Randomized Control Trials for Development? Three Problems*. https://www.brookings.edu/blog/future-development/2017/05/11/randomized-control-trials-for-development-three-problems/

HAMMER J., SPEARS D., 2016 – Village Sanitation and Child Health: Effects and External Validity in a Randomized Field Experiment in rural India. *Journal of Health Economics*, 48 : 135-148.

HANNAN E., 2008 – Randomized Clinical Trials and Observational Studies: Guidelines for Assessing Respective Strengths and Limitations. *JACC: Cardiovascular Interventions*, 2 (3) : 211-217.

HANNAN M. T., BRANDON N. T., 1990 – A Reassessment of the Effect of Income Maintenance on Marital Dissolution in the Seattle-Denver Experiment. *American Journal of Sociology*, 95 (5) : 1270-1298.

HARDIMAN D., 2000 – *Feeding the Baniya: Peasants and Usurers in Western India*. Oxford, Oxford University Press.

HARRISON G., 2011 – Randomization and Its Discontents. *Journal of African Economies*, 20 (4) : 626-652. https://doi.org/10.1093/jae/ejr030

HATHI P., HAQUE S., PANT L., COFFEY D., SPEARS D., 2017 – Place and Child Health: The Interaction of Population Density and Sanitation in Developing Countries. *Demography*, 54 (1) : 337-360.

HATT L., CHATTERJI M., MILES L., COMFORT A. B., BELLOWS B. W., 2014 – A False Dichotomy: RCTs and Their Contributions to Evidence-Based Public Health. *Global Health: Science and Practice*, 3 (1), 138-140.

HATT L., JOHNS B., CONNOR C., MELINE M., KUKLA M., MOAT K., 2015 – *Impact of Health Systems Strengthening on Health*. Bethesda, Health Finance and Governance Project, Abt Associates Inc.

HAUSMAN J. A., 1978 – Specification Tests in Econometrics. *Econometrica*, 46 (6) : 1251-1272.

HAUSMAN J. A., WISE D. A., 1985a – *Social Experimentation*. Chicago, University of Chicago Press.

HAUSMAN J. A., WISE D. A., 1985b – « Technical Problems in Social Experimentation: Cost Versus Ease of Analysis ». *In* HAUSMAN J. A., WISE D. A. (eds.) : *Social Experimentation*. Chicago, University of Chicago Press : 187-220.

HAUSMANN R., PRITCHETT L., RODRIK D., 2005 – Growth Accelerations. *Journal of Economic Growth*, 10 (4) : 303-329.

HAUSMANN R, KLINGER B., WAGNER R., 2008 – *Doing Growth Diagnostics in Practice: A 'Mindbook'*. CID Working paper, 177.

HAUSMANN R., RODRIK D., VELASCO A., 2008 – « Growth Diagnostics ». *In* SERRA N., STIGLITZ J. (eds) : *The Washington Consensus Reconsidered: Towards a New Global Governance*. Oxford, Oxford University Press : 324-355.

HECKMAN J. J., 1978 – Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46 (4) : 931-959.

HECKMAN J. J., 1990a – *Alternative Approaches to the Evaluation of Social Programs: Econometrics and Experimental Methods*. Conférence, Sixth World Meetings of the Econometric Society, Barcelone.

HECKMAN J. J., 1990b – Varieties of Selection Bias. *American Economic Review*, 80 (2) : 313-318.

HECKMAN J. J., 1992 – « Randomization and Social Policy Evaluation ». *In* MANSKI C. F., GARFINKEL I. (eds) : *Evaluating Welfare and Training Programs*. Cambridge, Harvard University Press : 201-230.

HECKMAN J. J., 2008 – Econometric Causality. *International Statistical Review*, 76 (1) : 1-27.

HECKMAN J. J., ASHENFELTER O., 1973 – « Estimating Labor Supply Functions ». *In* CAIN G. G., WATTS H. (eds) : *Income Maintenance and Labor Supply: Econometric Studies*. Chicago, Markham : 265-278.

HECKMAN J. J., HONORÉ B. E., 1990 – The Empirical Content of the Roy Model. *Econometrica*, 58 (5) : 1121-1149.

HECKMAN J. J., HOTZ J. V., 1989 – Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, 84 (408) : 862-874.

HECKMAN J. J., MOKTAN S., 2018 – *Publishing and Promotion in Economics: The Tyranny of the Top Five*. NBER Working Paper, 25093.

HECKMAN J. J., PINTO R., 2015 – Causal Analysis after Haavelmo, *Econometric Theory*, 31 (1) : 115-151.

HECKMAN J. J., PINTO R., 2019 – *Exploiting Noncompliance to Enhance Causal Inference of Randomized Controlled Trials*. Working Paper.

HECKMAN J. J., ROBB R., 1985 – « Alternative Methods for Evaluating the Impact of Interventions ». *In* HECKMAN J. J., SINGER B. S. (eds) : *Longitudinal Analysis of Labor Market Data*. New York, Cambridge University Press, 10 : 156-245.

HECKMAN J. J., ROBB R., 1986 – « Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes ». *In* WAINER H. (ed) : *Drawing Inferences from Self-Selected Samples*. New York, Springer-Verlag : 63-107.

HECKMAN J. J., SMITH J., 1995 – Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, 9 (2) : 85-110.

HECKMAN J. J., SMITH J., 1998 – « Evaluating the Welfare State ». *In* STROM S. (ed) : *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*. New York, Cambridge University Press : 241-318.

HECKMAN J. J., URZÚA S., 2010 – Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics*, 156 : 27-37.

HECKMAN J. J., VYTLACIL E., 2005 – Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73 (3) : 669-738.

HECKMAN J. J., VYTLACIL E., 2007 – « Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation ». *In* HECKMAN J. J., LEAMER E. (eds) : *Handbook of Econometrics.* Vol. 6B, Amsterdam, Elsevier : 4779-4874.

HECKMAN J. J., SMITH J. A., CLEMENTS N., 1997a – Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, 64 (4) : 487-535.

HECKMAN J. J., ICHIMURA H., TODD P. E., 1997b – Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies*, 64 (4) : 605-654.

HECKMAN J. J., LOCHNER L., TABER J. C., 1998a – Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents. *Review of Economic Dynamics*, 1 (1) : 1-58.

HECKMAN J. J., ICHIMURA H., SMITH J., TODD P. E., 1998b – Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66 (5) : 1017-1098.

HECKMAN J. J., LALONDE R. J., SMITH J. A., 1999 – « The Economics and Econometrics of Active Labor Market Programs ». *In* ASHENFELTER O. C., DAVID C. (eds) : *Handbook of Labor Economics*. Vol. 4A, New York, North-Holland : 1865-2097.

HECKMAN J. J., HOHMANN N., SMITH J., KHOO M., 2000 – Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. *Quarterly Journal of Economics*, 115 (2) : 651-694.

HECKMAN J. J., URZUA S., VYTLACIL E., 2006 – Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics*, 88 (3) : 389-432.

HEDT-GAUTHIER B. L., CHILENGI R., JACKSON E., MICHEL C., NAPUA M., ODHIAMBO J., BAWAH A., 2017 – Research Capacity Building Integrated into PHIT Projects: Leveraging Research and Research Funding to Build National Capacity. *BMC Health Services Research*, 17 (3) : 17-28.

HEMKENS L. G., CONTOPOULOS-IOANNIDIS D. G., Ioannidis J. P. A., 2016 – Agreement of Treatment Effects for Mortality from Routinely Collected Data and Subsequent Randomized Trials: Meta-epidemiological Survey. *British Medical Journal*, 352 : i493.

HERNÁN M. A., ROBINS J M., 2018 – *Causal Inference.* Boca Raton, Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

HES T., POLEDŇÁKOVÁ A., 2013 – Correction of the Claim for Microfinance Market of 1.5 Billion Clients. *International Letters of Social and Humanistic Sciences*, 2 (1) : 18-31.

HIBOU B., 2011 – *Anatomie politique de la domination*. Paris, La Découverte.

HIDALGO C., HAUSMANN R., 2009 – The Building Blocks of Economic Complexity. *PNAS*, 106 (26) : 10570-10575.

HINKELMANN K., KEMTHORNE O., 2008 – *Design and Analysis of Experiments.* New York, John Wiley.

HOFFMANN N., 2020 – Involuntary Experiments in Former Colonies: The Case for a Moratorium. *World Development*, 127 : 104805.

HOLLAND P. W., 1986 – Statistics and Causal Inference. *Journal of the American Statistical Association*, 81 (396) : 945-960.

HOPEWELL S., DUTTON S., YU L.-M., CHAN A.-W., ALTMAN D. G., 2010 – The Quality of Reports of Randomised Trials in 2000 and 2006: Comparative Study of Articles Indexed in PubMed. *British Medical Journal*, 340 : c723.

HORWITZ R. I., 1996 – The Dark Side of Evidence-Based Medicine. *Cleveland Clinic Journal of Medicine*, 63 (6) : 320-323.

HORWITZ R. I., SINGER B. H., 2017 – Why Evidence-Based Medicine Failed in Patient Care and Medicine-Based Evidence Will Succeed. *Journal of Clinical Epidemiology*, 84 : 14-17.

HOTZ V. J., 1992 – « Designing an Evaluation of the Job Training Partnership Act ». *In* MANSKI C., GARFINKEL I. (eds.) : *Evaluating Welfare and Training Programs*. Cambridge, Harvard University Press : 76-114.

HOUDE J.-F., JOHNSON T., LIPSCOMB M., SCHECHTER L., 2017 – *Pricing Winners: Optimizing Just-in-time Procurement Auctions in Dakar, Senegal*. Mimeo, University of Virginia.

HOUSE E. R., 2008 – Blowback: Consequences of Evaluation for Evaluation. *American Journal for Evaluation*, 29 (4) : 416-426.

HOUSE E. R., 2014 – « Origins of the Ideas in Evaluating with Validity ». *In* GRIFFITH J. C., MONTROSE-LOORHEAD B. (eds) : *Revisiting Truth, Beauty and Justice: Evaluating with Validity in the 21st Century, New Directions for Evaluation*. San Francisco, Jossey Bass, 142 : 9-10.

HUMMEL A., 2013 – « The Commercialization of Microcredits and Local Consumerism: Examples of Over-indebtedness from Indigenous Mexico ». *In* GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds) : *Microfinance, Debt and Over-indebtedness. Juggling with money*. Londres, Routledge : 253-271.

HUMPHREY J. H., 2009 – Child Undernutrition, Tropical Enteropathy, Toilets, and Handwashing. *The Lancet*, 374 (9694) : 1032-1035.

HUMPHREYS M., 2015 – *What Has Been Learned from the Deworming Replications: A Nonpartisan View.* www.columbia.edu/˜mh2245/w/worms.html

HUMPHREY J. H., MBUYA M. N. N., NTOZINI R., MOULTON L. H., STOLTZFUS R. J., TAVENGWA N. V., MUTASA K., MAJO F., MUTASA B., GOLDBERG M., 2019 – Independent and Combined Effects of Improved Water, Sanitation, and Hygiene, and Improved Complementary Feeding, on Child Stunting and Anaemia in Rural Zimbabwe: A Cluster-Randomised Trial. *The Lancet Global Health*, 7 (1) : e132-e147.

International Finance Corporation (IFC), 2017 – *Strategy and Business Outlook FY18-FY20. Creating Markets and Mobilising Private Capital*. Washington, International Finance Corporation.

IMBENS G., 2010 – Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48 (2) : 399-423.

IMBENS G., 2018 – Comments on Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright. *Social Science and Medicine*, 210 : 50-52.

IMBENS G., ANGRIST J., 1994 – Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62 (2) : 467-475.

Institute for Health Metrics and Evaluation (IHME), 2016 – *Financing Global Health 2016. Development Assistance, Public and Private Health Spending for the Pursuit of Universal Health Coverage*. Seattle, IHME.

Ioannidis J., 2005a – Why Most Published Research Findings are False. *PLoS Medicine*, 2 (8) : 1-6.

Ioannidis J., 2005b – Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *Journal of American Medical Association*, 294 (2) : 218-228.

Ioannidis J., 2018 – Randomized Controlled Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright. *Social Science & Medicine*, 210 : 53-56.

Ioannidis J., Haidich A.-B., Pappa M., Pantazis N., Kokori S. I., Tektonidou M. G., Contopoulos-Ioannidis D. G., Lau J., 2001 – Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies. *The Journal of the American Medical Association*, 286 (7) : 821-830.

Irwin D., 2019 – *Does Trade Reform Promote Economic Growth? A Review of Recent Evidence*. NBER Working Paper, 25927.

Jallais S., 2018 – D'un monde à l'autre *ou* les rhétoriques de l'exemple dans les manuels de micro-économie. *Revue de la régulation*, 23.

Jamison J. C., 2017 – *The Entry of Randomized Assignment into the Social Sciences*. Washington, World Bank.

Jamison D., Searle B., Galda K., Heyneman S. P., 1981 – Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement. *Journal of Educational Psychology*, 73 (4) : 556-567.

Jatteau A., 2013 – Expérimenter le Développement ? Des économistes et leurs terrains. *Genèses* 4 (93) : 8-28.

Jatteau A., 2016 – *Faire preuve par le chiffre ? Le cas des expérimentations aléatoires en* économie. Thèse de doctorat, Paris Saclay.

Jatteau A., 2018 – The Success of Randomized Controlled Trials: A Sociographical Study of the Rise of J-PAL to Scientific Excellence and Influence. *Historical Social Research/Historische Sozialforschung*, 43 (3[165]) : 94-119.

Javoy E., Rozas D., 2013 – « Estimating Levels of Credit Market Saturation ». *In* Guérin I., Labie M., Servet J.-M. (eds) : *The Crises of Microcredit*. Londres, Zed Books : 39-53.

Jayachandran S., de Laat J., Lambin E., Stanton C., Audy R., Thomas N., 2017 – Cash for Carbon: A Randomized Trial of Payments for Ecosystem Services to Reduce Deforestation. *Science*, 357 (6348) : 267-273.

Jensen R., 2010 – The (Perceived) Returns to Education and the Demand for Schooling. *Quarterly Journal of Economics*, 125 (2) : 515-548.

Jevons W. S., 1883 – *Methods of Social Reform*. Londres, Macmillan.

JINTARKANON S., NAKAPIEW S., TIENUDOM N., SUWANNAWONG P., WILSON D., 2005 – Unethical Clinical Trials in Thailand: A Community Response. *The Lancet*, 365 (9471) : 1617-1618.

JOHNSON S., ROGALY B., 1997 – *Microfinance and Poverty Reduction*. Londres, Oxfam.

JONES A., STEEL D., 2018 – A Combined Theoretical and Empirical Approach to Evidence Quality Evaluation: A Commentary on Deaton and Cartwright. *Social Science and Medicine*, 210 : 74-76.

JONES B., OLKEN B., 2008 – The Anatomy of Stop-Start Growth. *Review of Economics and Statistics*, 90 : 582-587.

JONSTON J., 1984 – *Econometric Methods*. New York, McGraw Hill.

JOSEPH N., 2013 – « Mortgaging Used Saree-skirts, Spear-heading Resistance: Narratives from the Microfinance Repayment Standoff in Ramanagaram, India, 2008-2010 ». *In* GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds) : *Microfinance, Debt and Over-indebtedness. Juggling with Money*. Londres, Routledge : 272-294.

J-PAL, 2013 – *Truth-telling in Third-Party Audits. J-PAL Policy Briefcase*. Cambridge, Abdul Latif Jameel Poverty Action Lab.

J-PAL, IPA, 2015 – Where Credit Is Due. *Policy Bulletin*, Cambridge, J-PAL/IPA.

KABEER N., 2019 – Randomized Control Trials and Qualitative Evaluations of a Multifaceted Programme for Women in Extreme Poverty: Empirical Findings and Methodological Reflections. *Journal of Human Development and Capabilities*, 20 (2) : 197-217.

KAFFENBERGER M., 2018 – Considering Construct Validity: Seemingly Minor Design Changes within the Same Project in Uganda Mait it Either the Best or Worst of all Global Literacy Interventions. *RISE*. https://riseprogramme.org/blog/considering_construct_validity

KANT I., 1998 [1785] – *Groundwork of the Metaphysics of Morals*. Cambridge, Cambridge University Press.

KAPLAN R., IRVIN V., 2015 – Likelihood of Null Effects in Large NHLBI Clinical Trials Has Increased over Time. *PLoS One*, 210 (8) : e0132382.

KAPPAGODA S., IOANNIDIS J., 2014 – Prevention and Control of Neglected Tropical Diseases: Overview of Randomized Trials, Systematic Reviews and Meta-analyses. *Bulletin of the World Health Organization*, 92 (5) : 356-366C.

KAPUR D., 2018 – Academic Research on India in the US: For Whom does the Bell Toll? *India in Transition*. https://theprint.in/opinion/academic-research-on-india-in-the-us-for-whom-does-the-bell-toll/78520/

KARING A., 2018 – *Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone*. Working Paper. https://drive.google.com/file/d/1Gq59ismP9V6I2pUzuLriMVC5t6y2MqX-/view

KARLAN D., APPEL J., 2011 – *More than Good Intentions: How a New Economics Is Helping to Solve Global Poverty*. New York, Dutton.

KARLAN D., MORDUCH J., 2017 – *Economics*. New York, McGraw Hill.

KARLAN D., ZINMAN J., 2009 – Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *The Review of Financial Studies*, 23 (1) : 433-464.

KARLAN D., ZINMAN J., 2011 – Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science*, 332 (6035) : 1278-1284.

KARLAN D., ZINMAN J., 2019 – Long-Run Price Elasticities of Demand for Credit: Evidence from a Countrywide Field Experiment in Mexico. *Review of Economic Studies*, 86 (4) : 1704-1746.

KASS N., 2001 – An Ethics Framework for Public Health. *American Journal of Public Health*, 91 (11) : 1776-1782.

KASY M., 2016 – Why Experimenters Might Not Always Want to Randomize, and What They Should Do Instead. *Political Analysis*, 24 : 324-338.

KASY M., SAUTMANN A., 2019 – *Adaptive Treatment Assignment in Experiments for Policy Choice*. Harvard/MIT, Preliminary draft, 2 juin.

KATES R. W., 2011 – What kind of a science is sustainability science? *Proceedings of the National Academy of Sciences*, 108 (49) : 19449-19450. https://doi.org/10.1073/pnas.1116097108

KAUFMANN D., MEHREZ G., T. GURGUR, 2002 – *Voice or Public Sector Management? An Empirical Investigation of Determinants of Public Sector Performance based on a Survey of Public Officials.* World Bank Research Working Paper.

KEANE M., 2010 – Structural vs. Atheoretic Approaches to Econometrics. *Journal of Econometrics*, 156 (1) : 3-20.

KEATING J., 2014 – Random Acts. What Happens when you Approach Global Poverty as a Science Experiment? *Slate*, 26 mars. http://www.slate.com/articles/business/crosspollination/2014/03/randomized_controlled_trials_do_they_work_for_economic_development.html

KELAHER M., NG L., KNIGHT K., RAHADI A., 2016 – Equity in Global Health Research in the New Millennium: Trends in First-Authorship for Randomized Controlled Trials among Low and Middle-Income Country Researchers 1990-2013. *International Journal of Epidemiology*, 45 (6) : 2174-2183.

KENNY C., PRITCHETT L., 2013 – *Promoting Millennium Development Ideals: The Risks of Defining Development Down*. Working Paper, Washington, Center for Global Development.

KERWIN J., THORNTON R. L., 2018 – *Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures*. Working Paper. https://doi.org/10.2139/ssrn.3002723

KHANDKER S. R., SAMAD H. A., KHAN Z. H., 1998 – Income and Employment Effects of Micro-credit Programmes: Village-level Evidence from Bangladesh. *The Journal of Development Studies*, 35 (2) : 96-124.

KIDD S., 2019 – The Demise of Mexico's Prospera Programme: A Tragedy Foretold. *Development Pathways*, 5 novembre. https://www.developmentpathways.co.uk/blog/the-demise-of-mexicos-prospera-programme-a-tragedy-foretold/

KINGI H., VILHUBER L., HERBERT S., STANCHI F., 2018 – *The Reproducibility of Economics Research: A Case Study*. Berkeley, BITSS Annual Meeting.

KLINE P., WALTERS C., 2016 – Evaluating Public Programs with Close Substitutes: The Case of Head Start. *Quarterly Journal of Economics*, 131 (4) : 1795-1848.

KOETSENRUIJTER W., 2017 – « Numbers in the News: More Ethos than Logos? ». *In* NGUYEN A. (ed.) : *News, Numbers and Public Opinion in a Data-Driven World*. New York, Bloomsbury : 260-276.

KOHL-ARENAS E., 2016 – *The Self-Help Myth: How Philanthropy Fails to Alleviate Poverty*. Berkeley, University of California Press.

KRAAY A., 2006 – When Is Growth Pro-poor? *Journal of Development Economics*, 80 : 198-227.

KRAMER M. S., SHAPIRO S. H., 1984 – Scientific Challenges in the Application of Randomized Trials. *JAMA: The Journal of the American Medical Association*, 252 (19) : 2739-2745.

KREMER M., 2003 – Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons. *American Economic Review (Papers and Proceedings)*, 93 (2) : 102-106.

KRISHNARATNE S., HENSEN B., CORDES J., ENSTONE J., HARGREAVES J. R., 2016 – Interventions to Strengthen the HIV Prevention Cascade: A Systematic Review of Reviews. *The Lancet HIV* 3 (7) : e307-e317.

KRUK M. E., YAMEY G., ANGELL S. Y., BEITH A., COTLEAR D., GUANAIS F., JACOBS L., SAXENIAN H., VICTORA C., GOOSBY E., 2016 – Transforming Global Health by Improving the Science of Scale-Up. *PLOS Biology*, 14 (3) : e1002360.

KUHN T. S., 1962 – *The Structure of Scientific Revolutions*. Chicago, The University of Chicago Press.

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. Capitalisme, institutions, pouvoirs*, 7. http://regulation.revues.org/7818

LABROUSSE A., 2016 – Not by Technique Alone. Comparing Development Analysis with Elinor Ostrom, Esther Duflo. *Journal of Institutional Economics*, 12 (2) : 277-303.

LABROUSSE A., 2017 – Learning from Randomized Controlled Experiments. The Narrative of Scientificity, Practical Complications, Historical Experience. *Books and Ideas*. http://www.booksandideas.net/Learning-from-Randomized-Controlled-Experiments.html

LALANDE A., 1902-1923 – *Vocabulaire critique et technique de philosophie*. Paris, PUF.

LALONDE R., 1986 – Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76 (4) : 604-620.

LAMBA S., SPEARS D., 2013 – Caste, "Cleanliness" and Cash: Effects of Caste-based Political Reservations in Rajasthan on a Sanitation Prize. *Journal of Development Studies*, 49 (11) : 1592-1606.

Lancet, 2004 – The World Bank is Finally Embracing Science. *The Lancet*, 364 : 731-732.

LAMPEDUSA DI G. T., 1960 [1958] – *The Leopard*. Londres, Collins and Harvill Press.

LAPORTE C., 2015 – *L'évaluation, un objet politique : le cas d'étude de l'aide au développement*. Thèse de doctorat, Institut d'études politiques de Paris.

LATOUR B., 2012 – *Enquête sur les modes d'existence. Une anthropologie des Modernes*. Paris, La Découverte.

Laura and John Arnold Foundation, 2018 – *Request for Proposals: Randomized Controlled Trials to Evaluate Social Programs Whole Delivery Will be Funded by Government or Other Entities*. Houston, Laura and John Arnold Foundation.

LAUTIER B., 2004 – *L'économie informelle dans le tiers-monde*. Paris, La Découverte.

LEE J., MORDUCH J., RAVINDRAN S., SHONCHOY A., ZAMAN H., 2021 – Poverty and Migration in the Digital Age: Experimental Evidence on Mobile Banking in Bangladesh. *American Economic Journal: Applied Economics*, 13 (1), 38-71.

LEE N. R., ROTHSCHILD M. L., W. SMITH, 2011 – Social Marketing Defined. *Social Marketing Quaterly*.

LEEUW F., VAESSEN J., 2009 – *Impact Evaluations and Development: NONIE Guidance on Impact Evaluation.* Washington, World Bank.

LEGOVINI A., DI MARO V., PIZA C., 2015 – *Impact Evaluation Helps Deliver Development Projects*. World Bank Working Paper, 7157.

LEIGH A., 2018 – *Randomistas. How Radical Researchers Changed Our World.* New Haven, Yale University Press.

LEMIEUX C., 2007 – À quoi sert l'analyse des controverses ? *Mil neuf cent. Revue d'histoire intellectuelle*, 1 : 191-212.

LENSINK R., 2014 – What Can We Learn from Impact Evaluations? *European Journal of Development Research*, 26 (1) : 12-17.

LEVINE R., 2006 – Some Recent Developments in the International Guidelines on the Ethics of Research Involving Human Subjects. *Annals of the New York Academy of Science*, 918 : 170-178.

LEVY S., 2006 – *Progress against Poverty: Sustaining Mexico's Progresa-Oportunidades Program.* Washington, Brookings Institution.

LEWIS A. W., 1954 – Economic Development with Unlimited Supplies of Labor. *Manchester School*, 22 : 139-191.

LILFORD R. J., JACKSON J., 1995 – Equipoise and the Ethics of Randomization. *Journal of the Royal Society of Medicine*, 88 (10) : 552-559.

LIST J., LUCKING-REILEY D., 2002 – The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign. *Journal of Political Economy*, 110 (1) : 215-233.

LIST J., RASUL I., 2011 – « Field Experiments in Labor Economics ». *In* Ashenfelter O., Card D. (eds) : *Handbook of Labor Economics*. Vol.4A, New York, North-Holland :103-228.

LITTLE I., MIRRLEES J., 1974 – *Project Appraisal and Planning for Developing Countries*. New York, Basic Books.

LONDON A. J., 2017 – Equipoise in Research. Integrating Ethics and Science in Human Research. *JAMA Guide to Statistics and Methods*, 317 (5) : 525-526.

LUBY S. P., RAHMAN M., ARNOLD B. F., UNICOMB L., ASHRAF S., WINCH P. J., STEWART C. P., BEGUM F., HUSSAIN F., BENJAMIN-CHUNG J., 2018 – Effects of Water Quality, Sanitation, Handwashing, and Nutritional Interventions on Diarrhoea and Child Growth in Rural Bangladesh: A Cluster Randomised Controlled Trial. *The Lancet Global Health*, 6 (3) : e302-e315.

LUCAS R., 1976 – A Critique on Econometric Policy Evaluation. *The Philips Curve and Labor Markets*, Carnegie-Rochester Conference on Public Policy, 1 : 19-46. https://doi.org/10.1016/S0167-2231(76)80003-6

LUCAS R. E., 1988 – On the Mechanics of Economic Development. *Journal of Monetary Economics*, 22 (1) : 3-42.

LUCAS R. E., SARGENT T. J., 1981 – *Rational Expectations and Econometric Practice*. Minneapolis, University of Minnesota Press.

LUNDBERG S., STEARNS J., 2019 – Women in Economics: Stalled Progress. *Journal of Economic Perspectives*, 33 (1) : 3-22.

LURIE P., WOLFE S. M., 1997 – Unethical Trials of Interventions to Reduce Perinatal Transmission of the Human Immunodeficiency Virus in Developing Countries. *New England Journal of Medicine*, 337 (5) : 853-856.

MAHJABEEN R., 2008 – Microfinancing in Bangladesh: Impact on Households, Consumption and Welfare. *Journal of Policy Modeling*, 30 (6) : 1083-1092.

MARSCHAK J., 1953 – « Economic Measurements for Policy and Prediction ». *In* HOOD W. C., KOOPMANS T. C. (eds) : *Studies in Econometric Method*. New Haven, Yale University Press : 1-26.

MAURER K., PYTKOWSKA J., 2014 – *Indebtedness of Microcredit Clients in Bosnia and Herzegovina*. Frankfurt, European Fund for Southeast Europe.

MACKENZIE D. A., MUNIESA F., SIU L., 2007 – *Do Economists Make Markets? On the Performativity of Economics*. Princeton, Princeton University Press.

MACLEOD M. R., MICHIE S., ROBERTS I., DIRNAGL U., CHALMERS I., IOANNIDIS J. P. A., AL-SHAHI SALMAN R. CHAN A.-W., GLASZIOU P., 2014 – Biomedical Research: Increasing Value, Reducing Waste. *The Lancet*, 383 (9912) : 101-104.

MacMurray J. J. V., 2010 – Systolic Heart Failure, *New England Journal of Medicine*, 362 : 228-238.

Mäki U., 1995 – Diagnosing McCloskey. *Journal of Economic Literature*, 33 (3) : 1300-1318.

Manski C. F., Garfinkel I., 1992 – *Evaluating Welfare and Training Programs*. Cambridge/Londres, Harvard University Press.

Manzi A., Mugunga J. C., Nyirazinyoye L., Iyer H. S., Hedt-Gauthier B., Hirschhorn L. R., Ntaganira J., 2018a – Cost-effectiveness of a Mentorship and Quality Improvement Intervention to Enhance the Quality of Antenatal Care at Rural Health Centers in Rwanda. *International Journal for Quality in Health Care*, 31 (5) : 359-364.

Manzi A., Nyirazinyoye L., Ntaganira J., Magge H., Bigirimana E., Mukanzabikeshimana L., Hirschhorn L. R., Hedt-Gauthier B., 2018b – Beyond Coverage: Improving the Quality of Antenatal Care Delivery through Integrated Mentorship and Quality Improvement at Health Centers in Rural Rwanda. *BMC Health Services Research*, 18 (1) : 1-8.

Martinez-Alonso E., Ramos J. M., 2016 – A Systematic Review of Randomized Clinical Trials Published in Malaria Journal between 2008 and 2013. *Rev. Esp. Quimioter*, 29 (3) : 130-145.

McCloskey D., 1983 – The Rhetoric of Economics. *Journal of Economic Literature*, 21 (2) : 481-517.

McKenzie D., 2012 – Beyond Baseline and Follow-up: The Case for More T in Experiments. *Journal of development Economics*, 99 (2) : 210-221.

McKenzie D., 2013 – How Should We Understand "Clinical Equipoise" When Doing RCTs in Development? World Bank. https://blogs.worldbank.org/impactevaluations/how-should-we-understand-clinical-equipoise-when-doing-rcts-development

McKenzie D., 2016 – Have RCTs Taken Over Development Economics? World Bank. https://blogs.worldbank.org/impactevaluations/have-rcts-taken-over-development-economics

McKenzie D., 2018 – Six Questions with Mark Rosenzweig. World Bank. https://blogs.worldbank.org/impactevaluations/six-questions-mark-rosenzweig

McKenzie D., 2019 – « Discussant's Comments ». *In* Basu K., Rosenblatt D., Sepulveda C. P. (eds) : *State of Economics, State of the World*. Cambridge, MIT Press : 488-493.

Meager R., 2019 – Understanding the Average Impact of Microcredit Expansion: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11 (1) : 57-91.

Meessen B., Hercot D., Noirhomme M., Ridde V., Tibouti A., Tashobya C. K., Gilson L., 2011 – Removing User Fees in the Health Sector: A Review of Policy Processes in Six Sub-Saharan African Countries. *Health Policy and Planning*, 26 (Suppl. 2) : ii16-ii29.

MEIER G. M., SEERS D. (eds.), 1984 – *Pioneers in Development*. Oxford, Oxford University Press.

MEYER M., HECK P., HOLTZMAN G., ANDERSON S., CAI W., WATTS D., CHABRIS C., 2019 – Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments. *PNAS*, 116 (22) : 10723-10728.

MIGUEL E., KREMER M., 2004 – Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72 (1) : 159-217.

MILES S. H., 2005 – *The Hippocratic Oath and the Ethics of Medicine*. Oxford, Oxford University Press.

MILFORD C., WASSENAAR D., SLACK C., 2006 – Resources and Needs of Research Ethics Committees in Africa: Preparations for HIV Vaccine Trials. *IRB: Ethics & Human Research*, 28 (2) : 1-9.

MILL J. S., 1843 – *A System of Logic, Ratiocinative and Deductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Evidence*. Londres, John Parker. https://www.gutenberg.org/files/27942/27942-pdf.pdf

MILLER F. G., BRODY H., 2007 – Clinical Equipoise and the Incoherence of Research Ethics. *Journal of Medicine and Philosophy*, 32 (2) : 151-165.

MILLER F. G., JOFFE S., 2011 – Equipoise and the Dilemma of Randomized Clinical Trials. *New England Journal of Medicine*, 364 (5) : 476-480.

MITCHELL S., GELMAN A., ROSS R., CHEN J., BARI S., HUYNH U. K. HARRIS M. W., EHRLICH SACHS S., STUART E., FELLER A. A., 2018 – The Millennium Villages Project: A Retrospective, Observational, Endline Evaluation. *The Lancet Global Health*, 6 (5) : e500--e513. https://doi.org/10.1016/S2214-109X(18)30065-2

MOFFITT R., 2004 – The Role of Randomized Field Trials in Social Science Research. *American Behavioral Scientist*, 47 (5) : 506-540.

MOFFITT R., 2006 – « Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective ». *In* SCHNEIDER B., MCDONALD S.-K. (eds) : *Scale-Up in Education: Ideas in Principle*. Lanham, Rowman and Littlefield : 173-186.

MONTE M., 2007 – L'oxymore : figure syntactico-sémantique ou élément d'une stratégie para-doxique? *Fabula*. https://www.fabula.org/atelier.php?L%27oxymore%3A_%26eacute%3Bl%26eacute%3Bment_d%27une_strat%26eacute%3Bgie_para%2Ddoxique%3F

MORGAN K. L., RUBIN D. B., 2012 – Rerandomization to Improve Covariate Balance in Experiments. *Annals of Statistics*, 40 (2) : 1263-1282.

MORDUCH J., 1999 – The Microfinance Promise. *Journal of Economic Literature*, 37 (4) : 1569-1614.

MORDUCH J., 2000 – The Microfinance Schism. *World Development*, 28 (4) : 617-629.

MORDUCH J., 2020 – Why RCTs Failed to Answer the Biggest Questions about Microcredit Impact. *World Development*, 127 : 104818.

MORDUCH J., SCHNEIDER R., 2017 – *The Financial Diaries*: *How American Families Cope in a World of Uncertainty*. Princeton, Princeton University Press.

MORVANT-ROUX S. (ed.), 2009 – *Exclusion et liens financiers : microfinance pour l'agriculture des pays du Sud*. Paris, Economica.

MORVANT-ROUX S., 2013 – « International Migration and Over-indebtedness in Rural Mexico ». *In* GUÉRIN I., MORVANT-ROUX S., VILLARREAL M. (eds) : *Microfinance, Debt and Over-Indebtedness: Juggling with Money.* Londres/ New York, Routledge : 170-192.

MORVANT-ROUX S., ROESCH M., 2015 – « The Social Credibility of Microcredit in Morocco after the Default Crisis ». *In* GUÉRIN I., LABIE M., SERVET J.-M. (eds) : *The Crises of Microcredit*. Londres, Zed Book : 113-130.

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 : 302-312.

MOSSE D., 2004 – *Cultivating Development: An Ethnography of Aid Policy and Practice.* Londres, Pluto Press.

MUDUR G., 2005 – India Plans to Audit Clinical Trials. *British Medical Journal*, 331 (7524) : 1044.

MUELLER U., 2019 – *A More Robust t-Test* (précédemment présenté sous le titre *Inference to the Mean*), présentation à Princeton. https://www.princeton.edu/~umueller/heavymean_slides.pdf

MÜLLER O., DE ALLEGRI M., BECHER H., TIENDREBOGO J., BEIERSMANN C., YE M., KOUYATE B., SIE A., JAHN A., 2008 – Distribution Systems of Insecticide-Treated Bed Nets for Malaria Control in Rural Burkina Faso: Cluster-Randomized Controlled Trial. *PLoS ONE*, 3 (9) : e3182.

MULLIGAN C., 2014 – The Economics of Randomized Experiments. *Economix Blog, New York Times*, 5 mars.

MUMMOLO J., PETERSON E., 2019 – Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review*, 113 (2) : 517-529.

MURALIDHARAN K., NIEHAUS P., SUKHTANKAR S., 2016 – Building State Capacity: Evidence from Biometric Smartcards in India. *American Economic Review*, 106 (10) : 2895-2929.

MURALIDHARAN K., NIEHAUS P., SUKHTANKAR S., WEAVER J., 2018a – *Improving Last-Mile Service Delivery Using Phone-Based Monitoring*. NBER Working Paper, 25298.

MURALIDHARAN K., NIEHAUS P., SUKHTANKAR S., WEAVER J., 2018b – Use Mobiles to Improve Governance. *Hindustan Times*, 5 décembre.

MURALIDHARAN K., NIEHAUS P., SUKHTANKAR S., 2018c – *General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India*. NBER Working Paper, 23838.

MURGAI R., RAVALLION M., VAN DE WALLE D., 2015 – Is Workfare Cost Effective against Poverty in a Poor Labor-Surplus Economy? *World Bank Economic Review*, 30 (3) : 413-445.

MUSGRAVE R., 2008 – « Merit Goods ». *In* DURLAUF S. N., BLUME L. E. (eds): *The New Palgrave Dictionary of Economics*. Londres, Palgrave Macmillan, 1 (8) : 4173-4176.

NADEL S., PRITCHETT L., 2016 – *Searching for the Devil in the Details: Learning about Development Program Design*. Center for Global Development Working Paper, 434.

National Bioethics Advisory Commission (NBAC), 2001 – *Ethical and Policy Issues in Research Involving Human Participants*. Bethesda, M. D., 1.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979 – *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, U.S. Department of Health, Education and Welfare.

NARITA Y., 2018 – *Toward an Ethical Experiment*. Yale University, Cowles Foundation Discussion Paper, 2127.

NAROTZKY S., BESNIER N., 2014 – Crisis, Value, and Hope: Rethinking the Economy: An Introduction to Supplement 9. *Current Anthropology*, 55 (S9) : S4-16.

NAUDET J.-D., 1999 – *Trouver des problèmes aux solutions. Vingt ans d'aide au Sahel*. Paris, OECD Éditions.

NAUDET J.-D., 2006 – Les OMD et l'aide de cinquième génération. *Afrique contemporaine*, 2 : 141-174.

NIEHAUS P., 2019 – RCTs: Why Scale Matters. *VoxDev video*. https://youtu.be/fD6MgGM5jWI

NIES A. S., EVANS G. H., SHAND D. G., 1973 – Regional Hemodynamic Effects of Beta-Adrenergic Blockade with Propranolol in the Unanesthetized Primate. *American Heart Journal*, 85 : 97-102.

NULL C., STEWART C. P., PICKERING A. J., DENTZ H. N., ARNOLD B. F., ARNOLD C. D., BENJAMIN-CHUNG J., CLASEN T., DEWEY K. G., FERNALD L. C. H., 2018 – Effects of Water Quality, Sanitation, Handwashing, and Nutritional Interventions on Diarrhoea and Child Growth in Rural Kenya: A Cluster-Randomised Controlled Trial. *The Lancet Global Health*, 6 (3) : e316-e329.

ODHIAMBO J., AMOROSO C. L., BAREBWANUWE P., WARUGABA C., HEDT-GAUTHIER B. L., 2017 – Adapting Operational Research Training to the Rwandan Context: Intermediate Operational Research Training Programme. *Global Health Action*, 10 (1). https://www.tandfonline.com/doi/full/10.1080/16549716.2017.1386930

OGDEN T. N., 2017 – *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. Cambridge, MIT Press.

OLIVIER DE SARDAN J.-P., 1995 – *Anthropologie et développement : essai en socio-anthropologie du changement social*. Paris, Karthala.

OLIVIER DE SARDAN J.-P., 2021 – *La revanche des contextes : des mésaventures de l'ingénierie sociale, en Afrique et au-delà*. Paris, Karthala.

OPEM L. C., GORONJA N., 2013 – *Responsible Finance: Reducing Over-indebtedness for Bosnia and Herzegovina's Microfinance Borrowers*. Washington, International Finance Corporation.

ORCUTT G. H., ORCUTT A. G., 1968 – Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes. *American Economic Review*, 58 (4) : 754-772.

ORR R., PANG N., PELLEGRINO E., SIEGLER M., 1997 – Use of the Hippocratic Oath: A Review of Twentieth Century Practice and a Content Analysis of Oaths Administered in Medical Schools in the U.S. and Canada in 1993. *Journal of Clinical Ethics*, 8 (4) : 377-388.

ÖZLER B., 2018 – *Incorporating Participants Welfare and Ethics into RCTs*. Washington, World Bank.

PALCA J., 1989 – AIDS Drug Trials Enter New Age. *Science, New Series*, 246 (4926) : 19-21.

PAMIÈS-SUMNER S., 2015 – *Development Impact Evaluation, State of Play and New Challenges*. Paris, AFD.

PARFIT D., 2011 – *On What Matters.* Oxford, Oxford University Press.

PATIL S. R., ARNOLD B. F., SALVATORE A. L., BRICENO B., GANGULY S., COLFORD Jr J. M., GERTLER P. J., 2014 – The Effect of India's Total Sanitation Campaign on Defecation Behaviors and Child Health in Rural Madhya Pradesh: A Cluster Randomized Controlled Trial. *PLoS medicine*, 11 (8) : e1001709.

PEARL J., 2009 – *Causality: Models, Reasoning and Inference*. New York, Cambridge University Press.

PEARL J., MACKENZIE D., 2018 – *The Book of Why. The New Science of Cause and Effect.* New York, Basic Books.

PETERS J., LANGBEIN J., ROBERTS G., 2018 – Generalization in the Tropics. Development Policy, Randomized Controlled Trials, and External Validity. *World Bank Research Observer*, 33 (1) : 34-64.

PETRYNA A., 2007 – Clinical Trials Offshored: On Private Sector Science and Public Health. *BioSocieties*, 2 (1) : 21-40.

PETRYNA A., 2009 – *When Experiments Travel: Clinical Trials and the Global Search for Human Subjects.* Princeton, Princeton University Press.

PETTICREW M., MCKEE M., LOCK K., GREEN J., PHILLIPS G., 2013 – In Search of Social Equipoise. *British Medical Journal*, 347.

PHILIBERT A., RAVIT M., RIDDE V., DOSSA I., BONNET E., BÉDECARRATS F., DUMONT A., 2017 – Maternal and Neonatal Health Impact of Obstetrical Risk

Insurance Scheme in Mauritania: A Quasi Experimental Before-and-After Study. *Health Policy and Planning*, 32 (3) : 405-417.

Picciotto J., 2011 – *Labors of Innocence*. Cambridge, Harvard University Press.

Picciotto R., 2012 – Experimentalism and Development Evaluation: Will the Bubble Burst? *Evaluation*, 18 (2) : 213-229.

Picciotto R., 2013 – Evaluation Independence in Organizations. *Journal of Multi-Disciplinary Evaluation*, 9 (20).

Picherit D., 2018 – Rural Youth and Circulating Labour in South India: The Tortuous Paths towards Respect for Madigas. *Journal of Agrarian Change*, 18 (1) : 178-195.

Pickering A. J., Null C., Winch P. J., Mangwadu G., Arnold B. F., Prendergast A. J., Njenga S. M., Rahman M., Ntozini R., Benjamin-Chung J., Stewart C. P., 2019 – The WASH Benefits and SHINE trials: interpretation of WASH intervention effects on linear growth and diarrhoea. *The Lancet Global Health*, 7 (8) : e1139-e1146.

Pinkovskiy M., Sala-I-Martin X., 2016 – Lights, Camera… Income! Illuminating the National Accounts-Household Surveys Debate. *The Quarterly Journal of Economics*, 131 : 579-631.

Pitt M. M., Khandker S. R., 1998 – The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter? *Journal of Political Economy*, 106 (5) : 958-996.

Plsek P. E., Greenhalgh T., 2001 – Complexity Science: The Challenge of Complexity in Health Care. *British Medical Journal*, 323 (7313) : 625-628.

Podsakoff P. M., MacKenzie S. B., Lee J. Y., Podsakoff N. P., 2003 – Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, 88 (5) : 879.

Porter T., 1995 – *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton, Princeton University Press.

Pradhan M., Suryahadi A., Sumarto S., Pritchett L., 2001 – Eating Like Which "Joneses"? An Iterative Solution to the Choice of a Poverty Line "Reference Group". *Review of Income and Wealth*, 47 (4) : 473-487.

Prasad V., Jorgenson J., Ioannidis J. P. A., Cifu A., 2013 – Observational Studies Often Make Clinical Practice Recommendations: An Empirical Evaluation of Authors' Attitudes. *Journal of Clinical Epidemiology*, 66 (4) : 361-366.

Prathap V., Khaitan R., 2016 – *When Is Microcredit Unsuitable? Guidelines Using Primary Evidence from Low-Income Households in India*. IFMR Finance Foundation Working Paper, WP-2016-01.

Pritchett L., 2000 – Understanding Patterns of Economic Growth: Searching for Hills among Mountains, Plateaus, and Plains. *World Bank Economic Review*, 14 : 221-250.

PRITCHETT L., 2005 – *The Political Economy of Targeted Safety Nets*. Washington, World Bank, Social Protection Discussion Paper Series, 501.

PRITCHETT L., 2006 – Who Is Not Poor? Dreaming of a World Truly Free of Poverty. *The World Bank Research Observer*, 21 : 1-23.

PRITCHETT L., 2010a – Is Microfinance a Shumpeterian Dead End? *Center for Global Development Blog*, 10 mars. http://blogs.cgdev.org/open_book/2010/05/is-microfinance-a-schumpeterian-dead-end.php

PRITCHETT L. (ed.), 2010b – *The Policy Irrelevance of the Economics of Education: Is Normative as Positive Useless or Worse*. Washington, Brookings Press.

PRITCHETT L., 2013a – *The Dangerous Seduction of the Kinky*. Cambridge, Center for International Development.

PRITCHETT L., 2013b – RCTs in Development, Lessons from the Hype Cycle. *Center for Global Development Blog*, 14 novembre. https://www.cgdev.org/blog/rcts-development-lessons-hype-cycle

PRITCHETT L., 2014a – Is Your Impact Evaluation Asking Questions that Matter? *Center for Global Devleopment Blog*. https://www.cgdev.org/blog/your-impact-evaluation-asking-questions-matter-four-part-smell-test

PRITCHETT L., 2014b – A Development Agenda without Developing Countries? The Politics of Penurious Poverty Lines (Part I). *Center for Global Development Blog*, 4 septembre. https://www.cgdev.org/blog/development-agenda-without-developing-countries-politics-penurious-poverty-lines-part-i

PRITCHETT L., 2014c – An Homage to the Randomistas on the Occasion of the J-PAL 10th Anniversary: Development as a Faith-Based Activity. *Center for Global Development Blog*, 10 mars. https://www.cgdev.org/blog/homage-randomistas-occasion-j-pal-10th-anniversary-development-faith-based-activity

PRITCHETT L., 2015 – Can Rich Countries Be Reliable Partners for National Development? *Horizons: Journal of International Relations and Sustainable Development*, 2 : 206-223.

PRITCHETT L., 2016 – Turns out Development Does Bring Development. *Center for Global Development Blog*, 21 septembre. https://www.cgdev.org/blog/turns-out-development-does-bring-development

PRITCHETT L., 2017 – The Perils of Partial Attribution: Let's All Play for Team Development. *Center for Global Development Blog*, 26 octobre. https://www.cgdev.org/publication/perils-partial-attribution

PRITCHETT L., 2018a – *The Debate about RCTs in Development Is Over: We Won. They Lost*. Conférence, New York, DRI.

PRITCHETT L., 2018b – Knowledge or Its Adoption? *Center for Global Development Blog*, 6 août. https://www.cgdev.org/publication/knowledge-or-its-adoption

PRITCHETT L., 2018c – We Knew Fire Was Hot. *RISE Blog*. https://www.riseprogramme.org/publications/we-knew-fire-was-hot

PRITCHETT L., J. SANDEFUR, 2013a – Claims to External Validity and Development Practice Don't Mix: Theory, Simulation, and Empirics. *Journal of Globalization and Development*, 4 : 161-197.

PRITCHETT L., SANDEFUR J., 2013b – *Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix*. Center for Global Development Working Paper, 336.

PRITCHETT L., SANDEFUR J., 2015 – Learning from Experiments when Context Matters. *American Economic Review: Papers and Proceedings*, 105 (5) : 471-475.

PRITCHETT L., SUMMERS L. H., 2014 – *Asiaphoria Meets Regression to the Mean*. NBER Working Paper, 20573.

PRITCHETT L., SAMIJ S., HAMMER J., 2012 – *It's All about MeE: Using Structured Experiential Learning ('e') to Crawl the Design Space*. HKS Faculty, Document de tavail.

PRITCHETT L., SAMIJ S., HAMMER J., 2013 – *It's All about MeE: Learning in Development Projects through Monitoring ('M'), Experiential Learning ('e') and Impact Evaluation ('E')*. Center for Global Development Working Paper, 233.

PRITCHETT L., SEN S., KAR S., RAIHANDE S., 2016 – Trillions Gained and Lost: Estimating the Magnitudes of Growth Episodes. *Economic Modelling*, 55 : 279-291.

PULLA P., 2018 – Link between Sanitation, Stunting Questioned, *The Hindu*. 3 février.

PUTNAM H., 2009 – *Renewing Philosophy*. Harvard, Harvard University Press.

QUENTIN A., GUÉRIN I., 2013 – La randomisation à l'épreuve du terrain. *Revue Tiers Monde*, 1 : 179-200.

RABIN M., 1993 – Incorporating Fairness into Game Theory and Economics. *American Economic Review*, 83 (5) : 1281-1302.

RAI A., SJÖSTRÖM T., 2004 – Is Grameen Lending Efficient? Repayment Incentives and Insurance in Village Economies. *Review of Economic Studies*, 71 (1) : 217-234.

RAMEY C. T., COLLIER A. M., SPARLING J. J., LODA F. A., CAMPBELL F. A., INGRAM D. A., FINKELSTEIN N. W., 1976 – « The Carolina Abecedarian Project: A Longitudinal and Multidisciplinary Approach to the Prevention of Developmental Retardation ». *In* THEODORE T. (ed.) : *Intervention Strategies for High-Risk Infants and Young Children*. Baltimore, University Park Press : 629-655.

RAO V., 2001 – Celebrations as Social Investments: Festival Expenditures, Unit Price Variation and Social Status in Rural India. *Journal of Development Studies*, 38 (1) : 71-97.

RAVALLION M., 2009a – Should the Randomistas Rule?. *Economists' Voice*, 6 (2) : 1-5.

RAVALLION M., 2009b – Evaluation in the Practice of Development, *World Bank Research Observer*, 24 (1) : 29-54.

RAVALLION M., 2012 – Fighting Poverty one Experiment at a Time: A Review Essay on Abhijit Banerjee and Esther Duflo. *Poor Economics, Journal of Economic Literature*, 50 (1) : 103-114.

RAVALLION M., 2014 – On the Implications of Essential Heterogeneity for Estimating Causal Impacts Using Social Experiments. *Journal of Econometric Methods*, 4 (1) : 145-151.

RAVALLION M., 2016 – *The Economics of Poverty: History, Measurement and Policy*. New York, Oxford University Press.

RAVALLION M., VAN DE WALLE D., DUTTA P., MURGAI R., 2015 – Empowering Poor People through Public Information? Lessons from a Movie in Rural India. *Journal of Public Economics*, 132 (December) : 13-22.

RAWLINS M., 2008 – De Testimonio: On the Evidence for Decisions about the Use of Therapeutic Interventions. *The Lancet*, 372 (9656) : 2152-2161.

RAYNAUD D., 2018 – *Sociologie des controverses scientifiques*. Paris, Éditions Matériologiques.

REDDY S., LAHOTY R., 2016 – $1.90 a day: What Does it Say? The New International Poverty Line. *New Left Review*, 97 : 106-127.

REDFIELD P., 2012 – Bioexpectations: Life Technologies as Humanitarian Goods. *Public Culture* 24 (1[66]) : 157-184.

REIDY W. J., RABKIN M., SYOWAI M., SCHAAF A., EL-SADR W. M., 2018 – Patient-level and Program-level Monitoring and Evaluation of Differentiated Service Delivery for HIV: A Pragmatic and Parsimonious Approach Is Needed. *AIDS (London, England)*, 32 (3) : 399-401.

REQUEJO J. H., BRYCE J., BARROS A. J. D., BERMAN P., BHUTTA Z., CHOPRA M., DAELMANS B., DE FRANCISCO A., LAWN J., MALIQI B., 2015 – Countdown to 2015 and Beyond: Fulfilling the Health Agenda for Women and Children. *The Lancet*, 385 (9966) : 466-476.

REVEL J. (ed.), 1996 – *Jeux d'échelles. La micro-analyse à l'expérience*. Paris, Gallimard/Le Seuil.

RIDDE V., HADDAD S., 2009 – Abolishing User Fees in Africa. *PLoS Medicine*, 6 (1) : e1000008.

RIOUX R., 2019 – *Réconciliations*. Paris, Débats Publics Éditions.

RODRICK D., 2008a – *One Economics, Many Recipes, Globalization, Institutions, and Economic Growth*. Princeton, Princeton University Press.

RODRIK D., 2008b – *Normalizing Industrial Policy*. Washington, World Bank, Commission on Growth and Development, Working Paper, 3.

RODRIK D., 2009 – « The New Development Economics: We Shall Experiment, but How Shall We Learn? ». *In* COHEN J., EASTERLY W. (eds) : *What Works in Development? Thinking Big and Thinking Small*. Washington, Brookings Institution Press : 24-47.

ROETHLISBERGER F. J., DICKSON W. J., 1939 – *Management and the Worker*. Cambridge, Harvard University Press, 5.

ROODMAN D., MORDUCH J., 2014 – The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence. *Journal of Development Studies*, 50 (4) : 583-604.

ROSENBAUM P., RUBIN D., 1983 – The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70 : 41-55.

ROSENBERG R., 2009 – *The New Moneylenders: Are the Poor Being Exploited by High Microcredit Interest Rates?* Washington, CGAP Occasional Paper, 15.

ROSENBOOM J. W., BAN R., 2017 – From New Evidence to Better Practice: Finding the Sanitation Sweet Spot. *Waterlines*, 36 (4) : 267-383.

Royal Academy of Sciences, 2019 – *The Prize in Economic Sciences 2019*. Stockholm, communiqué de presse. https://www.nobelprize.org/uploads/2019/10/press-economicsciences2019-2.pdf

Royal Swedish Academy of Sciences, 2019 – *Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019. Understanding Development and Poverty Alleviation*. Stockholm, Royal Swedish Academy of Sciences.

ROZAS D., 2014 – Microfinance in Mexico: Beyond the Brink. *European Microfinance Platform Blog*, 6 juin. http://www.e-mfp.eu/blog/microfinance-mexico-beyond-brink

RUBIN D. B., 1978 – Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6 (1) : 34-58.

RUHM C. J., 2019 – Shackling the Identification Police? *Southern Economic Journal*, 85 (4) : 1016-1026.

RUSSELL B., 1912 – *The Problems of Philosophy*. New York, Henry Holt and Co.

RUTTER H., SAVONA N., GLONTI K., BIBBY J., CUMMINS S., FINEGOOD D., GREAVES F., HARPER L., HAWE P., MOORE L., PETTICREW M., REHFUESS E., SHIELL A., THOMAS J., WHITE M., 2017 – The Need for a Complex Systems Model of Evidence for Public Health. *The Lancet*, 390 (10112) : 2602-2604.

SABET S. M., BROWN A., 2018 – Is Impact Evaluation Still on the Rise? The New Trends 2010–2015. *Journal of Development Effectiveness*, 10 (3) : 291-304.

SACHS J., 2001 – *Macroeconomics and Health: Investing in Health for Economic Development. Report of the Commission on Macroeconomics and Health*. Genève, World Health Organization.

SACHS J., 2005 – *The End of Poverty: Economic Possibilities for Our Time.* New York, Penguin.

SAMII C., 2016 – Causal Empiricism in Quantitative Research. *Journal of Politics*, 78 (3) : 941-955.

SANDEFUR J., 2015 – The Final Word on Microcredit? *Center for Global Development Blog*, 22 janvier. https://www.cgdev.org/blog/final-word-microcredit

SANSON-FISHER R. W., BONEVSKI B., GREEN L. W., D'ESTE C., 2007 – Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. *American Journal of Preventive Medicine*, 33 (2) : 155-161.

SARIN A., 2019 – Indecent Proposals in Economics. *The India Forum*. https://www.theindiaforum.in/article/indecent-proposals-economics

SATHYAMALA C., 2019 – In the Name of Science: Ethical Violations in the ECHO Randomised Trial. *Global Public Health*. https://doi.org/10.1080/17441692.2019.1634118

SAVAGE L. J., 1962 – *The Foundations of Statistical Inference: A Discussion Opened by L.J. Savage at the Meeting of the Joint Statistics Seminar, Birkbeck and Imperial Colleges, in the University of London*. New York, Barnes and Noble.

SAVEDOFF W. D., LEVINE R., BIRDSALL N. (eds), 2006 – *When Will We Ever Learn? Improving Lives through Impact Evaluation*. Washington, Center for Global Development.

SCHAFER A., 1982 – The Ethics of the Randomized Clinical Trial. *New England Journal of Medicine*, 307 (12) : 719-724.

SCHICKS J., 2013 – The Definition and Causes of Microfinance Over-Indebtedness: A Customer Protection Point of View. *Oxford Development Studies*, 41 (1) : S95-116.

SCHICKS J., ROSENBERG R., 2011 – *Too Much Microcredit? A Survey of the Evidence on Over-Indebtedness*. CGAP Occasional Paper, 19.

SCHILBACH F., 2019 – Alcohol and Self-control: A Field Experiment in India. *American Economic Review*, 109 (4) : 1290-1322.

SCHULER S. R., LENZI R., BADAL S. H., NAZNEEN S., 2018 – Men's Perspectives on Women's Empowerment and Intimate Partner Violence in Rural Bangladesh. *Culture, Health & Sexuality*, 20 (1) : 113-127.

SCHURMAN R., 2018 – Micro(soft) Managing a "Green Revolution" for Africa: The New Donor Culture and International Agricultural Development. *World Development*, 112 : 180-192.

SEHON S. R., D., STANLEY E., 2003 – A Philosophical Analysis of the Evidence-Based Medicine Debate. *BMC Health Services Research*, 3 (1) : 14.

SHAFFER P., 2015 – Two Concepts of Causation: Implications for Poverty. *Development and Change*, 46 (1) : 148-166.

SHAHAR E., 1997 – A Popperian Perspective of the Term "Evidence-Based Medicine". *Journal of Evaluation in Clinical Practice*, 3 (2) : 109-116.

SCOTT J. C., 1977 – *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. New Haven, Yale University Press.

SCOTT J. C., 1998 – *Seeing Like a State: How Certain Schemes for Improving the Human Condition Have Failed*. New Haven, Yale University Press.

SCRIVEN M., 1991 – *Evaluation Thesaurus*. Newbury Park, Sage Publications.

SCRIVEN M., 2008 – A Summative Evaluation of RCT Methodology and an Alternative Approach to Causal Research. *Journal of Multidisciplinary Evaluation*, 5 (9), 11-24.

Second International Study of Infarct Survival Collaborative Group (ISIS-2), 1988 – Randomised Trial of Intravenous Streptokinase, Oral Aspirin, Both, or Neither among 17,187 Cases of Suspected Acute Myocardial Infarction. *The Lancet*, 2 : 349-360.

SERVET J.-M., 2006 – *Banquiers aux pieds nus*. Paris, Odile Jacob.

SERVET J.-M., 2011 – La crise du microcrédit en Andhra Pradesh (Inde). *Revue Tiers Monde*, 3 : 43-59.

SERVET J.-M., 2018 – *L'économie comportementale en question*. Paris, Fondation pour le Progrès de l'Homme.

SERVET J.-M., TINEL B., 2020 – The Behavioural and Neoliberal Foundations of Randomisations. *Strategic Change: Briefings in Entrepreneurial Finance*, 29 (3) : 293-299.

SHAND D. G., 1975 – Propranolol. *New England Journal of Medicine*, 293 : 280-284.

SHAW L. W., CHALMERS T. C., 1970 – Ethics in Collaborative Clinical Trials. *Annals of the New York Academy of Sciences*, 169 (2) : 487-495.

SHELTON J. D., 2014 – Evidence-based Public Health: Not Only Whether It Works, But How It Can Be Made to Work Practicably at Scale. *Global Health: Science and Practice*, 2 (3) : 253-258.

SILVERMAN S., 2009 – From Randomized Controlled Trials to Observational Studies. *American Journal of Medicine* 122 (2) : 114-120.

SILVEY S. D., 1980 – *Optimal Design: An Introduction to the Theory for Parameter Estimation*. New York, Chapman/Hall.

SINGH I., SQUIRE L., STRAUSS J. (eds.), 1986 – *Agricultural Household Models: Extensions, Applications, and Policy*. Baltimore, The Johns Hopkins University Press.

SKINNER Q., 2003 – *Visions of Politics: Regarding Methods*. Cambridge, Cambridge University Press.

SKOUFIAS E., PARKER S., 2001 – Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico. *Economía* 2 (1) : 45-86.

SNOW J., 1855 – *On the Mode of Communication of Cholera*. Londres, Churchill. http://www.med.mcgill.ca/epidemiology/hanley/c609/material/SnowCholera/On_the_mode_of_communication_of_cholera.pdf

South African Cochrane Centre, 2014 – *Evidence-based Interventions for Diagnosing, Preventing and Treating Tuberculosis*. Cape Town, South African Cochrane Centre.

SPEARS D., GHOSH A., CUMMING O., 2013 – Open Defecation and Childhood Stunting in India: An Ecological Analysis of New Data from 112 Districts. *PloS one*, 8 (9) : e73784.

SQUIRE L., 1989 – « Project Evaluation in Theory and Practice ». *In* CHENERY H., SRINIVASAN T. N. (eds) : *Handbook of Development Economics*. Vol. 2, Amsterdam, North-Holland : 1093-1137.

SQUIRE L., VAN DER TAK H., 1975 – *Economic Analysis of Projects*. Baltimore/Londres, Johns Hopkins University Press/World Bank.

STACKHOUSE M., 2007 – *God and Globalization: volume 4 : Globalization and Grace*. New York, Continuum.

STAIGER D., STOCK J., 1997 – Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65 (3) : 557-586.

STAKE R. E., 2010 – *Qualitative Research: How Things Work*. New York, Guilford Press.

Statistics Canada, 2010 – *Survey Methods and Practices*. Ottawa, Ministère de l'Industrie.

STENBERG K., HANSSEN O., TAN-TORRES EDEJER T., BERTRAM M., BRINDLEY C., MESHREKY A., ROSEN J. E., STOVER J., VERBOOM P., SANDERS R., 2017 – Financing Transformative Health Systems towards Achievement of the Health Sustainable Development Goals: A Model for Projected Resource Needs in 67 Low-income and Middle-income Countries. *The Lancet Global Health*, 5 (9) : e875-e887.

STIGLITZ J., 1986 – The New Development Economics. *World Development*, 14 (2) : 257-265.

STIGLITZ J., 2004 – The Post-Washington Consensus. *The Initiative for Policy Dialogue*.

STIGLITZ J., 2006 – *Making Globalization Work*. New York, W.W. Norton.

STIGLITZ J., WEISS A., 1981 – Credit Rationing in Markets with Imperfect Information. *American Economic Review*, 71 : 393-410.

STOCK P. L., 1993 – The Function of Anecdote in Teacher Research. *English Education*, 25 (3) : 172-187.

STRASSMAN D., POLANYI L., 1995 – « The Economist as Storyteller ». *In* KUIPER E., SAP J. (eds) : *Out of the Margin: Feminist Perspectives on Economics*. Londres, Routledge : 129-150.

STUDENT [GOSSET W. S.], 1938 – Comparison between Balanced and Random Arrangements of Field Plots. *Biometrika*, 29 (3/4) : 363-378.

SUPPES P., 1982 – Arguments for Randomizing. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1982 (2) : 464-475.

SVORENČIK A., 2015 – *The Experimental Turn in Economics: A History of Experimental Economics*. Dissertation Series, 29, Utrecht, Utrecht School of Economics.

TAROZZI A., MAHAJAN A., BLACKBURN B., KOPF D., KRISHNAN L., YOONG J., 2014 – Micro-loans, Insecticide-treated Bednets, and Malaria: Evidence from a Randomized Controlled Trial in Orissa, India. *American Economic Review*, 104 (7) : 1909-1941.

TAROZZI A., DESAI J., JOHNSON K., 2015 – The Impacts of Microcredit: Evidence from Ethiopia. *American Economic Journal: Applied Economics*, 7 (1) : 54-89.

TARP F., 2009 – *Aid Effectiveness*. Helsinki, United Nations University/WIDER.

TAVERNISE S., 2015 – Few Health System Studies Use Top Method, Report Says. *New York Times*, 12 février.

TAYLOR M., 2011 – "Freedom from Poverty Is Not for Free": Rural Development and the Microfinance Crisis in Andhra Pradesh, India. *Journal of Agrarian Change*, 11 (4) : 484-504.

TAYLOR K. M, MARGOLESE R. G., SOLSKOLNE C. L., 1984 – Physicians' Reasons for Not Entering Eligible Patients in a Randomized Trial of Surgery for Breast Cancer. *New England Journal of Medicine*, 310 : 1363-1367.

TAYLOR-ROBINSON D.C., MAAYAN N., SOARES-WEISER K., DONEGAN S., GARNER P., 2015 – Deworming Drugs for Soil-transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance. *The Cochrane Database of Systematic Reviews*, 2015 (7) : CD000371.

TEELE D. L. (ed.), 2014 – *Field Experiments and Their Critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences*. New Haven/Londres, Yale University Press.

TENDLER J., 1993 – *New Lessons from Old Projects: The Workings of Rural Development in Northeast Brazil. A World Bank Operations Evaluation Study*. Washington, World Bank.

THALER R. H., 2015 – *Misbehaving. The Making of Behavioral Economics*. New York, Penguin.

THEO T., 2009 – « Philosophical Concerns in Critical Psychology ». *In* FOX D., PRILLENTENSKY I., AUSTIN S. (eds) : *Critical Psychology: An Introduction*. Sage Publications, Thousand Oaks, 36-53.

THEROUX P., FRANKLIN D., ROSS J. Jr., KEMPER W. S., 1974 – Regional Myocardial Function during Acute Coronary Occlusion and Its Modification by Pharmacologic Agents in the Dog. *Circulation Research*, 35 : 896-908.

THOMSON D. R., AMOROSO C., ATWOOD A., BONDS M. H., CYAMATARE RWABUKWISI F., DROBAC P., FINNEGAN K. E., FARMER D. B., FARMER P. E., HABINSHUTI A., 2018 – Impact of a Health System Strengthening Intervention on Maternal and Child Health Outputs and Outcomes in Rural Rwanda 2005-2010. *BMJ Global Health*, 3 : e000674.

TINBERGEN J., 1956 – *Economic Policy: Principles and Design*. Amsterdam, North Holland.

TODD P., WOLPIN K., 2006 – Assessing the Impact of a School Subsidy Program in Mexico using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling. *American Economic Review*, 96 (5) : 1384-1417.

UBEL P. A., SILBERGLEIT R., 2011 – Behavioral Equipoise: A Way to Resolve Ethical Stalemates in Clinical Research. *American Journal of Bioethics*, 11 (2) : 1-8.

United Nations, 2015 – *The Millennium Development Goals Report*. New York, United Nations.

United Nations Development Program, 1999 – *Human Development Report*. New York/Oxford, Oxford University Press.

VAN DER MEULEN RODGERS Y., BEBBINGTON A., BOONE C., DELL'ANGELO J., PLATTEAU J.-P., AGRAWAL A., 2020 – Experimental Approaches in Development and Poverty Alleviation. *World Development*, 127 : 104807.

VASS M., 2010 – *Prevention of Functional Decline in Older People: The Danish Randomised Intervention Trial on Preventative Home Visits*. København, Museum Tusculanum.

VEATCH R. M., 2007 – The Irrelevance of Equipoise. *Journal of Medicine and Philosophy*, 32 (2) : 167-183.

VEDUNG E., 2010 – Four Waves of Evaluation Diffusion. *Evaluation*, 16 : 263-277.

VICTORA C. G., BLACK R. E., TIES BOERMA J., BRYCE J., 2011 – Measuring Impact in the Millennium Development Goal Era and Beyond: A New Approach to Large-scale Effectiveness Evaluations. *The Lancet*, 377 (9759) : 85-95.

VISWANATHAN S., SAITH R., CHAKRABORTY A., PURTY N., MALHOTRA N., SINGH P., MITRA P., PADMANABHAN V., DATTA S., HARRIS J., GIDWANI S., 2019 – *Improving households' attitudes and behaviours to increase toilet use (HABIT) in Bihar*. New Delhi, International Initiative for Impact Evaluation, 3rd Grantee Final Report.

VIVALT E., 2019 – Specification Searching and Significance Inflation across Time, Methods and Disciplines. *Oxford Bulletin of Economics and Statistics*, 81 (4) : 797-816.

VIVALT E., 2020 – How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association*, 18 (6) : 3045-3089. http://repositorio.minedu.gob.pe/handle/123456789/4623

VIVALT E., COVILLE A., 2016 – *How Do Policymakers Update?*, Non publié.

VRIEZE J. DE, 2018 – The Metawars. *Science*, 361 (6408) : 1184-1188.

DE **W**AAL A., 1997 – *Famine Crimes: Politics and the Disaster Relief Industry in Africa*. Melton, James Currey.

WEBBER S., PROUSE C., 2018 – The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development. *Economic Geography*, 94 (2) : 166-187.

WEBER M., 1958 – « Science and Vocation ». *In* GERTH H. H., WRIGHT MILLS C. (eds) : *Max Weber: Essays in Sociology*. New York, New York University Press : 129-156.

WEIJER C., GLASS K. C., SHAPIRO S. H., 2000 – Why Clinical Equipoise, and Not the Uncertainty Principle, Is the Moral Underpinning of the RCT. *British Medical Journal*, 321 : 756-758.

WHIDDEN C., KAYENTAO K., LIU J. X., LEE S., KEITA Y., DIAKITÉ D., KEITA A., DIARRA S., EDWARDS J., YEMBRICK A., HOLEMAN I., SAMAKÉ S., PLEA B., COUMARÉ M., JOHNSON A. D., 2018 – Improving Community Health Worker Performance by Using a Personalised Feedback Dashboard for Supervision: A Randomised Controlled Trial. *Journal of Global Health*, 8 (2) : 020418.

WHITE H., 2013 – An Introduction to the Use of Randomised Control Trials to Evaluate Development Interventions. *Journal of Development Effectiveness*, 5 (1) : 30-49.

WHITE H., 2014 – Ten Things that Can Go Wrong with Randomised Control Trials. *Evidence Matters Blog*, International Initiative for Impact Evaluation. https://www.3ieimpact.org/blogs/ten-things-can-go-wrong-randomised-controlled-trials

WHITE H., MASSET E., 2018 – The Rise of Impact Evaluations and Challenges which CEDIL Is to Address. *Journal of Development Effectiveness*, 10 (4) : 393-399. https://doi.org/10.1080/19439342.2018.1539387

WHITTLE D., 2011 – If Not Randomized Trials, Then What? *Pulling for the Underdog Blog*. https://www.denniswhittle.com/2011/06/randomized-trials-not-silver-bullet.html

WILKE A., HUMPHREYS M., 2019 – *Field Experiments, Theory and External Validity*. Working Paper. https://www.dropbox.com/s/47s52xv0frrnvml/20190703_Wilke_Humphreys.pdf?dl=0

WMA General Assembly., 2014 – *Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. Fortaleza, World Medical Association.

WOOLCOCK M., 2013 – Using Case Studies to Explore the External Validity of "Complex" Development Interventions. *Evaluation*, 19 (3) : 229-248.

World Bank, 2005 – *World Development Report. Economic Growth in the 1990s: Learning from a Decade of Reform*. Washington, World Bank.

World Bank, 2012 – *World Bank Group Impact Evaluations: Relevance and Effectiveness*. Washington, Independent Evaluation Group.

World Bank, 2015 – *World Development Report: Mind, Society, and Behavior*. Washington, IBRD/World Bank.

World Bank, 2016 – *Transforming Development through Impact Evaluation, I2i DIME Annual Report*. Washington, World Bank.

World Health Organization, 2010 – *Monitoring the Building Blocks of Health Systems: A Handbook of Indicators and their Measurement Strategies*. Genève, WHO.

World Health Organization/World Bank, 2017 – *Tracking Universal Health Coverage: 2017 Global Monitoring Report*. Genève, WHO/IBRD/World Bank.

WORRALL J., 2007 – Why There's No Cause to Randomize. *The British Journal for the Philosophy of Science*, 58 (3) : 451-488.

WRONG M., 2009 – *It's Our Turn to Eat: The Story of a Kenyan Whistleblower*. New York, Harper.

WU D., 1973 – Alternative Tests of Independence between Stochastic Regressors and Disturbances. *Econometrica*, 41 (4) : 733-750.

WYDICK B., 2016 – Microfinance on the Margin: Why Recent Impact Studies May Understate Average Treatment Effects. *Journal of Development Effectiveness*, 8 (2) : 257-265.

WYDICK B., 2018 – Review of Randomistas: How Radical Researchers Changed Our World, *Development Impact Blog, World Bank*. https://blogs.worldbank.org/impactevaluations/review-randomistas-how-radical-researchers-changed-our-world

YANG Y., 2019 – *The Open Secret of Development Economics*. Project Syndicate, 26 octobre. https://www.strategicstudyindia.com/2019/10/the-open-secret-of-development-economics.html

YOU D., HUG L., EJDEMYR S., IDELE P., HOGAN D., MATHERS C., GERLAND P., NEW J. R., ALKEMA L., 2015 – Global, Regional, and National Levels and Trends in Under-5 Mortality between 1990 and 2015, with Scenario-based Projections to 2030: A Systematic Analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet*, 386 (10010) : 2275-2286.

YOUNG A., 2019 – Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *Quarterly Journal of Economics*, 134 (2) : 557-598.

YOUNG J., HARRISON J. WHITE G., MAY J., SOLOMON M., 2004 – Developing Measures of Surgeons' Equipoise to Assess the Feasibility of Randomized Controlled Trials in Vascular Surgery, *Surgery*, 136 (5) : 1070-1076.

YUSUF S., PETO R., LEWIS J., COLLINS R., SLEIGHT P., 1985 – Beta Blockade During and After Myocardial Infarction: An Overview of the Randomized Trials. *Progress in Cardiovascular Diseases*, 27 (5) : 335-371.

ZILIAK S. T., 2014 – Balanced versus Randomized Field Experiments in Economics: Why W. S. Gosset aka "Student" Matters. *Review of Behavioral Economics*, 1 : 167-208.

ZILIAK S. T., TEATHER-POSADAS E. R., 2016 – « The Unprincipled Randomization Principle in Economics and Medicine ». *In* DEMARTINO G., MCCLOSKEY D. (eds) : *Oxford Handbook on Professional Economic Ethics*. New York/Oxford, Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199766635.001.0001/oxfordhb-9780199766635-e-44

Table des illustrations

Chapitre 1

Chapitre 2

Chapitre 4

Chapitre 5

Chapitre 6

Chapitre 7

Chapitre 8

Table des tableaux

Liste des sigles et acronymes

3ie : International Initiative for Impact Evaluation
ACE : African Center of Excellence
AEA : American Economic Association
AFD : Agence française de développement
APD : Aide publique au développement
ATE : *Average Treatment Effect*
BHAT : *Beta-Blocker Heart Attack Trial*
BDO : *Block Development Officer*
BoP : *Bottom of the Pyramid*
BRAC : Building Resources Across Communities
CCT : *Conditional Cash Transfers*
CEGA : Center for Effective Global Action
CGAP : Consultative Group to Assist the Poor
CIE : Comité institutionnel d'éthique
CMV : Chaîne de valeur mondiale
CRI : Centre de recherches interdisciplinaires
CSU : Couverture sanitaire universelle
DFID : Department for International Development
DIME : Development Impact Evaluation
EDS : Enquête démographique et de santé
EF : Enquête finale
EI : Enquête initiale
EQM : Erreur quadratique moyenne
FBP : Financement basé sur la performance
FSI : *Fragile States Index*

GIEC : Groupe d'experts intergouvernemental sur l'évolution du climat

GISSI : Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico

GSDR : *Global Sustainable Development Report*

IDFC : International Development Finance Club

IMF : Institution de microfinance

IPA : Innovations for Poverty Action

IRB : Institutional Review Board

IRD : Institut de recherche pour le développement

IHME : Institute for Health Metrics and Evaluation

INR : *Indian National Rupee*

INR : Impôt négatif sur le revenu

Insee : Institut national de la statistique et des études économiques

ITT : *Intention To Treat*

J-PAL : Abdul Latif Jameel Poverty Action Lab

JTPA : *Job Training Partnership Act*

LATE : *Local Average Treatment Effects*

MBF : Massachusetts Bail Fund

MCO : Moindres carrés ordinaires

MDRC : Manpower Demonstration Research Corporation

MIT : Massachussets Institute of Technology

MMA : Méthodes mixtes et autres approches

MTE : *Marginal Treatment Effect*

ND : *National Development*

NDI : *National Development Index*

NIH : National Institute of Health

OCDE : Organisation de coopération et de développement économiques

OEO : Office of Economic Opportunity

ODD : Objectifs de développement durable

OMD : Objectifs du millénaire pour le développement

OMS : Organisation mondiale de la santé

ONG : Organisation non gouvernementale

PacDev : Pacific Development Consortium

PAS : Programme d'ajustement structurel

PCB : *Pollution Control Boards*

PDIA : *Problem-Driven Iterative Adaptation*

Pisa : Programme international pour le suivi des acquis des élèves

PME : Petites et moyennes entreprises

PPA : Parité de pouvoir d'achat

PPP : Partenariats public-privé

QJE : *Quarterly Journal of Economics*

RCT : *Randomized Controlled Trials*

SBM : Swachh Bharat Mission

SBS-EM : Solvary Brussels Schools of Economics and Managment

SPI : *Social Progress Index*

SR|ND : *Sector-Wide Reforms|National Development*

SUTVA : *Stable Unit Treatment Value Assumption*

SWF : *Social Welfare Function*

TP (Y) : *Targeted programs (Income)*

TP (S) : *Targeted programs (Sectors)*

Unicef : United Nations International Children's Emergency Fund

UPS : Unité primaire de sondage

USAID : United States Agency for International Development

VAN : Valeur actualisée nette

VI : Variable instrumentale

VRO : *Village Revenue Officer*

WASH : *Water, Sanitation and Hygiene*

WASH-B : WASH-Benefits

Résumés

Chapitre 1. Les *randomistas* doivent-ils (continuer à) faire la loi ?

Martin Ravallion

La popularité croissante des expérimentations randomisées dans le domaine du développement s'accompagne de débats incessants sur les mérites de cette approche. Ce chapitre passe en revue les questions qu'elle soulève. Il explique en quoi le fait de privilégier systématiquement les *Randomized Controlled Trials* (RCT) est problématique, et ce, à trois égards. Premièrement, les raisons de cette préférence ne sont, *a priori*, pas claires. Par exemple, avec un budget donné, même une étude non expérimentale biaisée peut se rapprocher davantage de la vérité qu'une RCT coûteuse. Deuxièmement, les objections éthiques soulevées à l'encontre des RCT ne sont pas dûment prises en compte par leurs partisans. Troisièmement, privilégier systématiquement les RCT entraîne un biais de sélection de ce qui est évalué, ce qui risque de fausser les éléments de preuve nécessaires à l'élaboration des politiques. À l'avenir, les questions posées et la manière d'y répondre devraient être dictées par les lacunes les plus pressantes dans nos connaissances, et non par les préférences méthodologiques de certains chercheurs. Le véritable étalon-or est la meilleure méthode pour répondre à la question qui se présente.

Chapitre 2. Randomiser le développement : méthode ou folie pure ?

Lant Pritchett

Un argument important en faveur de l'utilisation accrue des méthodes d'analyse randomisée (RCT) dans le domaine du développement est que les résultats issus de ces études vont favoriser l'adoption de programmes et de projets efficaces (à la fois en freinant les projets inopérants et en améliorant la conception de nouveaux projets), permettant ainsi de réduire la pauvreté et d'améliorer le bien-être humain. Des preuves empiriques de différents pays font toutefois apparaître qu'une transformation quadridimensionnelle du développement

national – à savoir vers des économies plus productives, des États plus réactifs, des organisations et des administrations plus efficientes et un traitement social plus égalitaire – produit des bénéfices en termes de réduction de la pauvreté et de bien-être humain infiniment supérieurs aux meilleurs résultats que l'on peut espérer obtenir avec des programmes améliorés. Les arguments soutenant que la recherche basée sur les RCT est un bon (sans parler du « meilleur ») investissement reposent sur le fait de croire *à la fois* qu'il est très peu probable que des études qui ne s'appuient pas sur des RCT puissent faire progresser le développement national, *et* qu'il est très fortement probable que les enseignements tirés de RCT puissent améliorer les résultats.

Chapitre 3. Le pouvoir subversif des expérimentations aléatoires

Jonathan MORDUCH

Deux types de RCT très différents sont utilisés par les économistes, même s'ils sont souvent confondus. Le premier type est de nature évaluative, et sert à déterminer si une politique ou une action a eu un effet positif ou non. Les critiques pointent du doigt que le fait de privilégier ces RCT par rapport à d'autres méthodes d'évaluation peut limiter les connaissances. Le second type est la RCT exploratoire, qui interroge la façon dont les comportements, les institutions et les marchés réagissent à l'évolution des prix, des contrats et à d'autres facteurs économiques. En venant perturber le système économique actuel par l'intermédiaire d'un modèle expérimental, ces RCT exploratoires ouvrent de nouveaux horizons à la micro-économie empirique comme nulle autre méthode n'est capable de le faire. On peut être partagé entre l'envie de mettre les RCT évaluatives sur un piédestal et celle d'encourager dans le même temps les RTC exploratoires. Des exemples de RCT appliquées aux domaines de l'assurance, du microcrédit et de l'argent numérique viennent illustrer ces débats.

Chapitre 4. Les expérimentations aléatoires dans l'économie du développement, leurs détracteurs et leur évolution

Timothy OGDEN

L'utilisation des RCT dans l'économie du développement a suscité un flot constant de critiques, mais relativement peu de réactions de la part des *randomistas* (hormis un nombre croissant de praticiens et d'articles). Ce chapitre fait une synthèse des critiques et examine la difficulté d'y répondre directement. Par la suite, ce chapitre applique le cadre *Problem-Driven Iterative Adaptation* (PDIA) de Lant Pritchett, éminent détracteur des RCT, pour illustrer comment le mouvement des RCT a répondu aux critiques, sinon aux auteurs de ces critiques, par une évolution constante de ses pratiques. Enfin, ce chapitre évalue la situation présente du mouvement RCT en termes d'impact et de productivité.

Chapitre 5. Réduire le déficit des connaissances dans la prestation de soins de santé à l'échelle mondiale : apports et limites des expérimentations aléatoires

Andres Garchitorena, Megan B. Murray, Bethany Hedt-Gauthier, Paul E. Farmer et Matthew H. Bonds

Les RCT sont considérées comme l'étalon-or en matière d'évaluation d'impact dans le domaine du développement international. Elles sont associées à une nouvelle vague de politiques de santé globale basées sur des preuves empiriques. L'utilisation des RCT pour répondre à certaines questions fondamentales dans le domaine de la santé globale pose toutefois des questions : si des solutions sont connues, réalisables à l'échelle et étayées par les preuves existantes, pourquoi des centaines de millions de personnes n'ont-elles pas accès à des services de santé essentiels ? Le manque de clarté sur les méthodes de recherche adaptées au renforcement des systèmes de santé fait écho à un manque d'investissements plus généralisé dans des systèmes de prestation de soins intégrés, plus complexes et adaptatifs. Le présent chapitre étudie l'utilisation des RCT dans le domaine de la santé globale en soulignant leurs apports majeurs et en abordant certaines priorités urgentes en matière de recherche sur la mise en œuvre, à un moment où les Objectifs de développement durable (ODD) mettent l'accent sur l'importance des approches sectorielles, comme les soins de santé primaire intégrés et la couverture maladie universelle.

Chapitre 6. Essais et tribulations : l'essor et le déclin des expérimentations aléatoires dans le secteur de l'eau, de l'assainissement et de l'hygiène

Dean Spears, Radu Ban et Oliver Cumming

Ce chapitre présente les débats autour des expérimentations randomisées dans le secteur de l'eau, de l'assainissement et de l'hygiène (*Water, Sanitation and Hygiene* – WASH), et en tire des enseignements pour les politiques de développement en général. L'assainissement est un cas intéressant, d'une part parce que l'amélioration des systèmes sanitaires est largement reconnue comme un élément essentiel du processus de développement, d'autre part car les interventions dans le secteur WASH sont souvent moins adaptées aux expérimentations randomisées que d'autres sujets appartenant aux sciences de la santé ou à l'économie du développement. Le présent chapitre traite d'une série récente d'évaluations randomisées qui, loin de régler définitivement les questions importantes des politiques d'assainissement en milieu rural, ont ravivé la confusion et le débat dans ce domaine. En effet, même des interventions sanitaires parfaitement conçues et mises en œuvre peuvent produire des effets qui diffèrent de l'une à l'autre ou en fonction des différents contextes, et les faits et théories issus de sources autres que les RCT sont nécessaires (en complément de celles-ci) pour apporter des réponses complètes et opportunes aux problématiques des politiques d'assainissement. Enfin, nous argumentons en faveur d'un recours accru aux

RCT dans le secteur WASH sur un point qu'elles pourraient permettre d'éclairer en particulier : répondre à des questions concernant les comportements, plutôt que la santé elle-même.

Chapitre 7. Les expérimentations aléatoires en microfinance : miracle ou mirage ?

Florent Bédécarrats, Isabelle Guérin et François Roubaud

Le microcrédit constitue depuis longtemps un thème central pour les RCT dans le domaine du développement, dont le point d'orgue a été la publication d'un numéro spécial consacré à six RCT conduites dans différentes régions du monde dans une revue d'économie prestigieuse. Ce numéro spécial a été salué comme étant la première étude rigoureuse et, en théorie, irréfutable sur les impacts du microcrédit. Or, une analyse détaillée de la mise en œuvre de ces six RCT révèle de nombreuses limites en termes de validité interne et externe, d'éthique et d'interprétation. Ce chapitre emploie des outils analytiques proposés par la statistique, l'économie politique et l'anthropologie du développement pour déterminer dans quelle mesure l'ensemble de la chaîne de production des RCT s'écarte de ses principes de base (depuis l'échantillonnage, la collecte, la saisie et le recodage des données à la publication et à la diffusion des résultats en passant par les estimations et l'interprétation). Il questionne également le décalage entre le succès académique et politique de ce numéro spécial et les nombreuses incohérences dans la mise en œuvre des études qui le composent.

Chapitre 8. La supériorité rhétorique de *Poor Economics*

Agnès Labrousse

L'analyse textuelle met en évidence la rhétorique du livre *Poor Economics* et suggère qu'elle participe du succès déroutant des RCT en économie. Elle montre comment Banerjee et Duflo combinent efficacement (1) le *logos* (discours rationnel, utilisation extensive des chiffres), (2) le *pathos* (anecdotes percutantes visant à émouvoir le lecteur, chiffres personnifiés) et (3) l'*ethos* (les narrateurs font preuve de sagesse, d'excellence et de bonne volonté). Malgré leur rejet explicite des anecdotes, les auteurs en font un usage ubiquitaire. Cela devient moins paradoxal si l'on considère les multiples fonctions persuasives de ces petits récits personnifiés. Les anecdotes ont en outre un rôle heuristique discret. Ce chapitre examine également les effets persuasifs et épistémiques de deux schémas rhétoriques récurrents : le juste milieu entre deux extrêmes, d'une part, et la rhétorique des petites mesures produisant de grands effets, d'autre part, qui amplifie le micro et minimise le macro. Cette rhétorique habile, qui intègre des composantes manipulatoires, ne doit pas occulter le manque de rigueur de ces récits et l'étendue des angles morts des RCT révélés par le hors-discours.

Chapitre 9. Les *randomistas* sont-ils des évaluateurs ?

Robert Picciotto

Dans un article publié en 2012, l'auteur concluait que la vague d'enthousiasme suscitée par la randomisation était vouée à ne pas durer. Mais il avait sous-estimé l'attrait du public pour les RCT et leur adaptation aux évolutions des exigences émanant d'un marché de l'évaluation dominé par des intérêts particuliers. Il est désormais clair que la bulle de la randomisation n'éclatera pas de sitôt. Ancrées par des racines historiques profondes, défendues par les puissants de ce monde et jugées d'une grande rigueur par un public mal informé, les RCT continueront à être plébiscitées, en dépit de leurs limitations statistiques et éthiques, leur incapacité à traiter les questions de recherche sociale complexes et leur inefficacité en tant qu'outils de responsabilisation et d'apprentissage des organisations.

Chapitre 10. Expérimentations aléatoires et éthique : les économistes doivent-ils se soucier de l'*équipoise* ?

Michel Abramowicz et Ariane Szafarz

L'équipoise est définie par Freedman (1987) comme un « état de réelle incertitude de la part du chercheur clinicien concernant les mérites thérapeutiques relatifs de chaque branche de l'alternative ». Ce principe naît de l'injonction éthique selon laquelle, en cas de supériorité de l'une des branches, on lèse les patients qui recevraient l'autre traitement. En économie du développement, les expérimentations contrôlées randomisées RCT ignorent souvent l'exigence d'équipoise et tendent à désavantager le groupe-contrôle. Ce chapitre examine comment le principe d'équipoise est formalisé dans la littérature médicale et présente les arguments relatifs à son éventuelle prise en compte par les économistes. Il souligne que deux circonstances rendent l'équipoise particulièrement nécessaire : d'une part, lorsque l'expérimentation n'est pas effectuée en double, voire simple, aveugle et, d'autre part, lorsque le groupe contrôle comporte des personnes particulièrement vulnérables. Au-delà, ce chapitre plaide pour la mise en place d'un large débat sur l'éthique des RCT en sciences économiques et sociales.

Chapitre 11. Utilisation des *a priori* dans les protocoles expérimentaux : que laisse-t-on de côté en les ignorant ?

Eva Vivalt

Une grande question revient dans toutes les études académiques, celle de savoir si et comment il faut exploiter les croyances préalables (*a priori*) des décideurs dans les protocoles d'étude. Tant les RCT que les évaluations d'impact qui n'en sont pas pourraient éclairer plus efficacement l'élaboration des politiques si les *a priori* étaient pris en compte. Les « non-RCT » peuvent tirer profit de l'assignation déterministe pour maximiser la valeur de l'information qu'elles

apportent. Les non-RCT semblent aussi se prêter davantage à la recherche de spécifications, ce qui devrait conduire les décideurs à être plus sceptiques sur la façon d'interpréter leurs résultats. Par conséquent, ils devraient peut-être être plus convaincus par les preuves empiriques issues des RCT et se montrer désireux d'expérimenter davantage en tirant parti de leurs croyances préalables. Je discute de ces questions et j'utilise des simulations pour explorer les avantages potentiels de l'utilisation d'*a priori* dans la conception des études.

Chapitre 12. Épilogue : La randomisation et l'évaluation des politiques sociales revisitées

James J. HECKMAN

Ce chapitre examine le cas des expérimentations randomisées en économie. Il revient sur le précédent article de l'auteur, intitulé « *Randomization and Social Policy Evaluation* », et actualise son propos. Le chapitre présente un résumé de l'histoire de la randomisation en économie. Il identifie deux grandes vagues d'enthousiasme pour la méthode, surnommées les « Deux Réveils » (*Two Awakenings*) en raison du zèle quasi religieux associé à chaque vague. La première vague a largement contribué au développement de la micro-économétrie, de par la nature surprenante des preuves expérimentales. La seconde vague a amélioré les plans expérimentaux en vue d'éviter certains des problèmes statistiques techniques identifiés par les économètres dans le sillage de la première vague. Cependant, les questions conceptuelles profondes concernant les paramètres estimés, l'interprétation économique et la pertinence politique des résultats expérimentaux n'ont pas été abordées lors de cette seconde vague.

Contributeurs

Michel Abramowicz est cardiologue à l'hôpital Érasme, l'hôpital universitaire de l'université libre de Bruxelles (ULB), en Belgique, dont il a supervisé le séminaire pluridisciplinaire pendant quatre ans. Il a publié des articles et des lettres dans des revues scientifiques telles que l'*American Journal of Cardiology*, l'*American Journal of Respiratory and Critical Care Medicine*, l'*American Journal of Epidemiology*, l'*European Heart Journal* et le *New England Journal of Medicine*. Lorsqu'il était chef de l'unité de soins coronariens, il a participé activement, en tant qu'enquêteur de terrain, à la deuxième étude internationale sur la survie à l'infarctus (ISIS-2), l'une des premières méga-RCT en cardiologie.

Radu Ban est responsable du programme « Eau, assainissement et hygiène » à la fondation Bill & Melinda Gates. Il dirige les travaux de mesure et de preuve de ce programme. À ce titre, il gère un programme de recherche visant à mieux comprendre quelles approches marchent en faveur de l'assainissement et pourquoi. Radu Ban est économiste du développement de formation, et a obtenu son doctorat à la London School of Economics. Avant de rejoindre la fondation, il a travaillé comme économiste à la Banque mondiale dans le cadre de l'initiative Development Impact Evaluation (DIME) en se concentrant sur la gouvernance et le développement communautaire.

Florent Bédécarrats est titulaire d'un doctorat de l'université de Paris 1 Panthéon-Sorbonne. Depuis octobre 2022, il est chercheur à l'IRD, membre de l'Unité mixte internationale SOURCE et titulaire d'une Chaire de Professeur Junior pour le développement de méthodes d'évaluation des politiques et solutions d'adaptation au changement climatique. Il est aussi chercheur affilié à l'équipe DIAL (UMR LEDa). Il a été entre 2019 et 2022 responsable du management de la donnée à Nantes Métropole, de 2013 à 2019 en charge de la coordination des évaluations d'impact scientifique à l'AFD et de 2007 à 2013, responsable des activités de recherche et du développement de Cerise, une plateforme de soutien aux institutions de microfinance. Auparavant, il a travaillé pendant trois ans en Amérique latine dans une entreprise solidaire spécialisée dans le tourisme et la culture au Brésil, pour un réseau de coopératives de microfinance au Mexique ou encore pour une ONG internationale au Guatemala.

Matthew Bonds est professeur assistant à la Harvard Medical School. Il est titulaire d'un doctorat en économie et d'un doctorat en écologie de l'université

de Géorgie. Il a rejoint le corps professoral de la Harvard Medical School après un post-doctorat en développement durable sous la direction de Jeffrey Sachs à l'Earth Institute de l'université de Columbia. Tout en développant des cadres théoriques formels sur les trappes à pauvreté, il a travaillé avec Partners In Health au Rwanda, et il est le co-fondateur et co-directeur général de l'ONG de santé Pivot à Madagascar. Ses recherches portent sur l'élaboration de nouvelles méthodes d'évaluation en matière de santé mondiale.

Oliver Cumming est professeur assistant à la London School of Hygiene and Tropical Medicine et directeur adjoint du groupe de santé environnementale. Il est actuellement chercheur principal dans le cadre d'études menées dans plusieurs pays, dont le Mozambique, le Kenya, le Sénégal et la République démocratique du Congo, et occupe le poste de directeur de recherche au sein du consortium SHARE. Ses recherches se concentrent sur l'épidémiologie des maladies liées à l'eau, l'assainissement et l'hygiène (*Water, Sanitation and Hygiene* – WASH), et il dirige actuellement de multiples expérimentations pour évaluer l'effet de différentes interventions WASH sur divers aspects de la santé des enfants, notamment les infections entériques, la croissance et le développement des enfants, et les performances des vaccins oraux.

Sir Angus Deaton est professeur émérite à l'université de Princeton et professeur présidentiel à l'University of Southern California (USC). Il est l'auteur de *La grande évasion* et a co-écrit, avec Anne Case, *Morts de désespoir. L'avenir du capitalisme.* Il travaille sur la santé, le bonheur, le développement, la pauvreté, les inégalités et la question des preuves pour orienter les politiques publiques. Membre de l'Académie nationale des sciences, membre de l'Académie britannique et membre honoraire de la Royal Society of Edinburgh, il a reçu en 2015 le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel. Il est né à Édimbourg, en Écosse. Il a été fait *Knight Bachelor* en 2016.

Paul Farmer est décédé le 21 février 2022 à Kigali, au Rwanda. Docteur en médecine de l'université de Harvard, il était professeur et titulaire de la chaire Kolokotrones du département de santé mondiale et de médecine sociale à la Harvard Medical School. Il est co-fondateur et a été directeur de la stratégie de Partners in Health, une organisation internationale à but non lucratif qui, depuis 1987, fournit des services de soins de santé directs et entreprend des activités de recherche et de défense des intérêts des personnes malades et vivant dans la pauvreté. Il était professeur de médecine et chef de la division de l'équité en matière de santé mondiale au Brigham and Women's Hospital. Il était également conseiller spécial auprès du secrétaire général des Nations unies pour les leçons de la médecine communautaire en Haïti

Andres Garchitorena est chercheur à l'IRD. Il est titulaire d'un doctorat en médecine vétérinaire et en santé publique. Il a rejoint les rangs de l'IRD après un post-doctorat en santé mondiale à la Harvard Medical School. Il est également directeur scientifique adjoint de Pivot, une ONG fortement axée sur la recherche, qui a pour mission de renforcer le système de santé et qui travaille en partenariat avec le ministère de la Santé de Madagascar, où il réalise de nouvelles

évaluations sur l'impact des interventions de renforcement des systèmes de santé en utilisant des données sur la population et le système de santé.

Isabelle Guérin est titulaire d'un doctorat. Elle est socio-économiste, directrice de recherche à l'IRD, affiliée à l'Institut français de Pondichéry, et membre de la School of Social Sciences de l'Institute for Advanced Study, Princeton (2019-2020). Elle est spécialisée dans l'économie politique et morale de l'argent, de la dette et des finances. Ses travaux actuels se concentrent sur la financiarisation des économies nationales en examinant comment elle génère de nouvelles formes d'inégalités et de domination, mais aussi des initiatives alternatives et solidaires. Ses travaux se basent le plus souvent sur ses propres données originales recueillies sur le terrain, combinent l'ethnographie et l'analyse statistique, et s'inscrivent dans une perspective interdisciplinaire et comparative. Elle mène également une réflexion permanente sur les conditions de production des données et la combinaison des méthodes.

James Heckman est titulaire de la chaire Henry Schultz d'économie et de politiques publiques à l'université de Chicago. Dans le cadre de ses travaux, il cherche à comprendre les origines des inégalités et de la formation des compétences, et développe et applique des stratégies pour aborder ces questions. Heckman a publié plus de 300 articles et neuf livres. Il a reçu, entre autres, le prix de la Banque de Suède en sciences économiques en mémoire d'Alfred Nobel, le prix Dan David et le prix de l'Amitié du Gouvernement chinois. Il est directeur du Centre pour l'économie du développement humain à l'université de Chicago, qui étudie les sources de la pauvreté et de l'immobilité sociale, ainsi que les politiques visant à améliorer l'épanouissement humain.

Bethany Hedt-Gauthier est professeure associée à la Harvard Medical School. Biostatisticienne, elle est reconnue pour le développement, l'application et l'évaluation de méthodologies de recherche innovantes visant à améliorer la santé des populations vivant dans des environnements aux ressources limitées. Ses recherches actuelles portent principalement sur le renforcement des systèmes de santé en Afrique, et plus particulièrement sur la chirurgie dans le monde. En outre, elle dirige des travaux sur l'équité dans le cadre des collaborations de recherche en santé globale, en mettant notamment l'accent sur des programmes de formation complets visant à renforcer les capacités nationales en matière de soutien et de direction de la recherche.

Agnès Labrousse est maîtresse de conférences en économie à l'université de Picardie et rédactrice en chef adjointe de la *Revue de la régulation*. Elle a travaillé sur l'épistémologie, l'industrie pharmaceutique et les questions de développement dans une perspective institutionnaliste. Ses travaux sur les RCT font le lien entre ces domaines de recherche et son article « Not by Technique Alone. Comparing Development Analysis with Elinor Ostrom and Esther Duflo », paru dans le *Journal of Institutional Economics*, qui a reçu à la fois le prix EAEPE Kapp-Prize en 2016 et le prix Ostrom en 2017.

Jean-Paul Moatti est professeur émérite d'économie de la santé à Aix-Marseille université (AMU) et a beaucoup travaillé sur le sida, la tuberculose et le

paludisme, ainsi que sur l'accès aux médicaments essentiels dans les pays en développement. Entre mars 2015 et février 2020, il a été directeur général de l'IRD, en charge du partenariat scientifique avec les pays en développement. Il a été membre du groupe indépendant de scientifiques qui a rédigé le premier rapport d'évaluation quadriennal des Nations unies sur la mise en œuvre des objectifs de développement durable (*Global Sustainable Development Report* – GSDR 2019).

Jonathan Morduch est professeur de politique publique et d'économie à la Wagner Graduate School of Public Service de l'université de New York. Ses recherches portent sur la pauvreté, les inégalités et la finance. Il est l'auteur, avec Rachel Schneider, de *The Financial Diaries*: *How American Families Cope in a World of Uncertainty* (MORDUCH et SCHNEIDER, 2017) et co-auteur de *Portfolios of the Poor*: *How the World's Poor Live on $2 a Day* (COLLINS *et al.*, 2009). M. Morduch a également co-écrit *The Economics of Microfinance* (ARMENDÁRIZ et MORDUCH, 2010) et *Economics* (KARLAN et MORDUCH, 2017). Il est l'un des fondateurs et le directeur exécutif de la NYU Financial Access Initiative.

Megan Murray est professeure de santé globale et de médecine sociale, professeure associée de médecine au département de médecine globale et de santé mondiale de la Harvard Medical School, et professeure d'épidémiologie à la Harvard School of Public Health. Elle dirige le centre de recherche du département de santé globale et de médecine sociale ; elle est également directrice de la recherche au sein de la division de l'équité en santé mondiale du Brigham and Women's Hospital. Outre ses travaux de recherche remarquables en épidémiologie des maladies infectieuses, elle s'intéresse également aux évaluations des effets du renforcement des systèmes de santé au Rwanda et à Madagascar.

Gulzar Natarajan travaille dans la fonction publique indienne. Au cours de ses vingt ans de carrière, il a travaillé dans le cabinet du Premier ministre de l'Inde, il a dirigé la société d'infrastructure de l'État d'Andhra Pradesh, a été collecteur du district d'Hyderabad, président et directeur général d'une société de distribution d'électricité basée à Visakhapatnam, commissaire municipal de Vijayawada, et a occupé des postes sur le terrain dans tout l'Andhra Pradesh. Il a également dirigé la conception et la mise en œuvre de vastes projets dans les domaines des infrastructures, de l'urbanisme, de la santé, de l'éducation, des compétences et des moyens de subsistance, de la réduction de la pauvreté, etc. à différents niveaux du gouvernement. Il est titulaire d'une licence en ingénierie de l'Institut indien de technologie de Chennai et du Master in Public Administration in International Development (MPA-ID) de la Harvard Kennedy School.

Timothy N. Ogden est directeur général de la Financial Access Initiative à NYU-Wagner, et chercheur principal du programme d'opportunités économiques et du programme de sécurité financière de l'Institut Aspen. Il est également partenaire exécutif de Sona Partners, président du conseil d'administration de GiveWell, et président de la Fondation du syndrome de Bardet Biedl. Ogden est rédacteur en chef du *faiV*, un bulletin d'information très lu sur l'inclusion

financière, la finance numérique, les politiques fondées sur des preuves et le développement économique. Son livre, *Experimental Conversations: Perspectives on the Use of Randomized Trials in Development Economics* (OGDEN, 2017), rassemble des entretiens réalisés avec 20 éminents penseurs sur le sujet.

Ila Patnaik est professeure au National Institute of Public Finance and Policy à New Delhi. Auparavant, elle était conseillère économique principale auprès du gouvernement indien. Ses recherches portent sur la macro-économie internationale, la finance, les cycles économiques des économies émergentes et la réglementation du secteur financier. Elle a publié des articles dans des revues spécialisées telles que le *Journal of International Money and Finance*, *The World Bank Economic Review* et l'*International Finance*. Le Dr Patnaik a également fait partie de divers groupes de travail du ministère des Finances.

Robert Picciotto est professeur associé à l'université d'Auckland et conseiller principal indépendant en matière d'évaluation auprès du ministère des Affaires étrangères et du Commerce en Nouvelle-Zélande. Il est diplômé de l'université de Princeton et membre de l'Académie des sciences sociales. Il a pris sa retraite de la Banque mondiale en 2002 après avoir occupé plusieurs postes opérationnels et de gouvernance institutionnelle, notamment celui de vice-président, de responsable de la planification et du budget et celui de directeur général du groupe d'évaluation indépendant pendant deux mandats consécutifs de cinq ans.

Lant Pritchett est directeur de recherche du projet RISE à la Blavatnik School of Government d'Oxford et associé au projet *Building State Capability* à la Harvard Kennedy School. Après avoir obtenu son doctorat en économie en 1983 au Massachussets Institute of Technology (MIT), il a travaillé à la Banque mondiale de 1988 à 2007. Il a vécu en Indonésie de 1998 à 2000 et en Inde de 2004 à 2007. Il a enseigné à la Harvard Kennedy School entre 2000 et 2018. Il a plus d'une centaine de publications à son actif (avec plus de cinquante co-auteurs différents) sur un large éventail de sujets liés au développement (éducation, croissance économique, capacité de l'État, mobilité de la main-d'œuvre, pauvreté et apprentissage à partir des RCT).

Martin Ravallion est actuellement titulaire de la première chaire Edmond D. Villani d'économie à l'université de Georgetown. Avant de rejoindre Georgetown en 2013, il était directeur du Groupe de recherche sur le développement (département de recherche de la Banque mondiale). Il a rejoint la Banque mondiale en 1988 et a travaillé dans presque tous les secteurs et toutes les régions au cours des 24 années qui ont suivi. Avant de rejoindre la Banque mondiale, il était professeur à l'Australian National University. Au cours des 30 dernières années, il s'est principalement intéressé à la pauvreté et aux politiques de lutte contre celle-ci. Il a conseillé de nombreux gouvernements et agences internationales sur ce sujet.

Rémy Rioux est conseiller maître à la Cour des comptes. Expert auprès des institutions financières internationales, il a, au cours de sa carrière consacrée au développement et à l'Afrique, occupé des postes de haut niveau. Après avoir été directeur de cabinet du ministre de l'Économie, des Finances et du Commerce

extérieur, Pierre Moscovici, il a été nommé secrétaire général adjoint par Laurent Fabius, ministre des Affaires étrangères et du Développement international, et a coordonné l'agenda « finance » de la présidence française de la COP21. Depuis 2016, il dirige l'Agence française de développement (AFD). Rémy Rioux est également président de l'International Development Finance Club (IDFC), le premier fournisseur au monde de financements pour le développement et le climat.

François Roubaud est économiste et statisticien, directeur de recherche à l'IRD, membre de l'équipe DIAL (UMR LEDa, laboratoire d'économie de Dauphine) à Paris, dont il a été le directeur entre 2000 et 2004, et actuellement basé au Brésil (Université Fédérale Rio de Janeiro). Titulaire d'un doctorat en économie de l'université Paris-Ouest Nanterre, il est aussi diplômé de l'École nationale de la statistique et de l'administration économique (ENSAE) de Paris. En statistique, il a lancé l'approche des enquêtes mixtes (ménages-entreprises) pour mesurer l'économie informelle, notamment l'*enquête 1-2-3*, et a mis au point les modules de gouvernance greffés sur les enquêtes officielles auprès des ménages, désormais utilisés pour le suivi de l'Objectif de développement durable (ODD) 16. Ces deux méthodes sont reconnues comme des normes internationales et mises en œuvre dans des dizaines de pays (en Afrique, en Amérique latine et en Asie). Ses recherches sur l'économie du développement portent plus particulièrement sur le marché du travail et l'économie informelle, la corruption, la gouvernance, et les institutions, ainsi que sur l'évaluation de l'impact et l'économie politique des politiques de développement.

Dean Spears est directeur exécutif du Research Institute for Compassionate Economics (RICE). Il est démographe économique et économiste du développement. Ses recherches portent sur différents domaines, notamment la santé, la croissance et la survie des enfants, en particulier en Inde ; l'environnement, la pollution et le changement climatique ; et les dimensions démographiques du bien-être social. En plus d'être l'un des directeurs exécutifs fondateurs du RICE, il est professeur assistant d'économie à l'université du Texas à Austin, économiste invité à l'unité économique et de planification de l'Institut indien de statistique à Delhi. Il est également affilié à l'Initiative Climate Futures de l'Université de Princeton. Avec Diane Coffey, il est l'auteur du livre primé *Where India Goes*: *Abandoned Toilets, Stunted Development, and the Costs of Caste* (Coffey, Spears, 2017).

Ariane Szafarz est professeure de finance à l'ULB, à la Solvary Brussels Schools of Economics and Managment (SBS-EM), et co-directrice du Centre européen de recherche en microfinance (CERMi). Elle est titulaire d'un doctorat en mathématiques et d'un doctorat en philosophie des sciences. Elle travaille actuellement sur la microfinance, la banque sociale, leurs dérives de mission et la discrimination sexuelle. Elle a publié plusieurs livres et articles, notamment dans les revues *Academy of Management Review*, *European Economic Review*, *Journal of Banking and Finance*, *Journal of Business Ethics*, *Journal of Development Studies*, *Journal of International Money and Finance*, *Review*

of Finance, *World Development*. Deux des articles qu'elle a co-écrits ont reçu le prix Warren Samuels (en 2016 et 2019) décerné par l'Association for Social Economics lors des réunions de l'Allied Social Science Association (ASSA).

Eva Vivalt est maîtresse de conférences en économie à l'Australian National University. Elle s'intéresse principalement à l'étude des obstacles à la prise de décisions politiques fondées sur des données probantes, y compris les questions méthodologiques et la manière dont les données probantes sont interprétées et utilisées. Elle est également chercheuse principale de la RCT sur le revenu de base « Y Combinator Research » et s'intéresse également au développement, à l'économie comportementale et à l'altruisme efficace. Elle a récemment travaillé sur les prévisions de résultats en sciences sociales et a créé, avec Stefano DellaVigna, la Social Science Prediction Platform, une plateforme que les chercheurs peuvent utiliser pour obtenir des prévisions pour leurs propres études.

En 2019, le prix de la *Sveriges Riksbank en sciences économiques en Mémoire d'Alfred Nobel* était décerné à trois des principaux promoteurs des expérimentations aléatoires (*Randomized Controlled Trials* – RCT), pour leur contribution essentielle à la lutte contre la pauvreté. Utilisée de longue date en médecine, l'expérimentation devient aujourd'hui la référence dans le champ des politiques économiques et sociales. Cette suprématie est-elle scientifiquement légitime, éthiquement satisfaisante et politiquement désirable ? Les RCT ont-elles réellement « considérablement amélioré notre capacité à combattre la pauvreté », dont la pandémie de Covid-19 a encore accru l'urgence ? À quelles questions permettent-elles de répondre et quelles autres sont laissées de côté ? Cet ouvrage de référence réunit 26 spécialistes mondiaux sur ce thème (y compris deux éminents lauréats du prix Nobel d'économie), issus d'origines différentes et d'un large spectre de disciplines (économie, statistique, anthropologie, philosophie, santé publique, etc.). Il présente, dans un langage accessible à tous, les principales forces et faiblesses des RCT dans le champ du développement : comment elles fonctionnent, ce qu'on peut en attendre, pourquoi parfois elles échouent, comment elles pourraient être améliorées, et pourquoi d'autres méthodes sont à la fois utiles et nécessaires.

Florent Bédécarrats est chercheur à l'IRD et membre de l'UMI Source, où il occupe une chaire de professeur junior sur l'évaluation des politiques d'adaptation au changement climatique.

Isabelle Guérin est socio-économiste, directrice de recherche à l'IRD, membre de l'UMR Cessma et chercheure affiliée à l'Institut français de Pondichéry.

François Roubaud est économiste et statisticien, directeur de recherche à l'IRD, membre de l'UMR LEDa-DIAL et chercheur invité à l'UFRJ.

35 €

IRD Editions

AFD

www.editions.ird.fr

ISSN 2431-7128
ISBN 978-2-7099-2947-9