Version 2 (février 2017)

Article pour le GDR Économie & Sociologie

Faire preuve par le chiffre ?
Une approche « par le bas » des expérimentations aléatoires

Arthur Jatteau, IDHES (ENS Paris Saclay)
arthur.jatteau@ens-paris-saclay.fr

Plan

Introduction. Que prouvent les expérimentations aléatoires ?

Vers la fin des années 1990, quand les premières expérimentations aléatoires sont menées dans des pays en développement (Glewwe et al., 2004 ; Glewwe, Kremer et Moulin, 2009 ; Miguel et Kremer, 2004), elles sont en partie oubliées dans les pays occidentaux. Peu se souviennent des années 1960-1980 aux États-Unis, où des milliards de dollars étaient dépensés pour évaluer des projets expérimentaux autour de l'impôt négatif ou de l'aide au logement (Monnier, 1992). L'assignation aléatoire était alors considérée comme le *must* des techniques d'évaluation et était pratiquée à grande échelle. Mais dès les années 1980, sous l'administration Reagan, la méthode, si elle ne disparaît pas totalement, passe de mode, suite à un certain nombre de critiques sans réponse satisfaisante apportée par ses promoteurs (différences d'objectifs et de temporalité entre chercheurs et décideurs politiques, complexité de généraliser les résultats…). Ce n'est qu'à partir du début des années 2000, avec la création du laboratoire Poverty Action Lab en 2003 (devenu le J-PAL deux ans après) au prestigieux Massachussetts Institue of Technology (MIT), à Boston, que des chercheurs, regroupés autour de l'économiste Esther Duflo, vont réemployer massivement cette méthode d'évaluation d'impact essentiellement dans les pays en développement. Une quinzaine d'années après sa renaissance, ce sont autour de mille expérimentations aléatoires qui ont été menées dans le monde entier, dont plus de huit cents par

le seul J-PAL. En peu de temps, cette méthode s'est largement répandu et fait aujourd'hui partie des classiques de l'évaluation quantitative.

Mais de quoi s'agit-il exactement ? L'expérimentation aléatoire, que l'on rencontre également en français sous le terme d'évaluation (par assignation) aléatoire, désigne une méthode d'évaluation d'impact de programme à l'aide du tirage au sort, à la manière de ce qui se fait en médecine avec les essais cliniques randomisés. Prenons l'exemple d'une distribution de manuels scolaires à des écoliers. Quels effets ont-ils sur leur niveau scolaire ? Pour le savoir, on divise une population d'élèves en deux groupes. Un groupe, dit de « traitement », reçoit les manuels scolaires. Un autre, le groupe de contrôle, ne reçoit rien. Au bout d'une année scolaire, par exemple, on fait passer des tests à tous les élèves. En comparant les moyennes entre les deux groupes, on peut ainsi voir si les manuels scolaires ont un impact sur le niveau. Instinctivement, on voit bien que la validité de cette méthode repose sur la comparabilité des groupes. S'il a été décidé de distribuer les manuels uniquement aux écoliers les plus faibles, alors la comparaison des groupes ne permettra pas de dire quelque chose, en toute rigueur, de l'effet des manuels scolaires (dans ce cas, on parle de biais de sélection). Ainsi, pour peu que la population de départ soit suffisamment nombreuse, le tirage au sort permet de minimiser les chances d'avoir des différences significatives entre eux. La force apparente des expérimentations aléatoires, c'est qu'il n'y a aucune hypothèse à ajouter, à rebours de certaines approches économétriques.

Pour ses partisans, que l'on nomme également « randomisation » pour insister sur la centralité méthodologique du tirage aléatoire, elles constitueraient ainsi le modèle ultime de l'administration de la preuve. Elle seule permettrait en effet de fournir ce qu'Abhijit Banerjee, un économiste du MIT qui a fondé le J-PAL avec Esther Duflo et Sendhil Mullainathan, nomme de la « hard evidence » (Banerjee, 2007). Par la grâce du tirage au sort, qui lui seul permettrait de pallier au biais de sélection, on atteindrait un stade inégalé de confiance dans la mesure de ce que l'on souhaite tester. La randomisation se trouverait ainsi tout en haut de la pyramide des preuves (Laurent et al., 2009). À ce titre, les économistes pro-randomisation pratiquent avec succès un véritable travail de militantisme méthodologique, aussi bien auprès du monde académique que d'institutions.

Jusque récemment, le concert de louanges autour de cette méthode était saisissant et les voix critiques étaient peu nombreuses. Il est possible que les choses soient en train de changer quelque peu, notamment avec l'obtention du « Nobel » d'économie en 2015 par Angus Deaton, un économiste de Princeton farouche opposant à la randomisation (Deaton, 2010 ; Deaton et Cartwright, 2016). Avec Dani Rodrik, également chercheur à Princeton et Martin Ravallion, ancien directeur de la recherche de la Banque mondiale, ils sont, chez les économistes « mainstream », les critiques les plus acerbes de la méthode (Ravallion, 2009 ; Rodrik, 2008). Des critiques ont également vue le jour chez des économistes hétérodoxes comme chez des sociologues (Bardet et Cussó, 2012 ; Bédécarrats, Guérin et Roubaud, 2013 ; Quentin et Guérin, 2013 ; Reddy, 2012). L'article séminal en la matière est celui d'Agnès Labrousse (Labrousse, 2010), qui dresse un ensemble de critiques par la suite approfondies par d'autres.

La plupart de ces publications adopte une posture qu'on pourrait qualifier d'externaliste, en ce qu'elle considère la méthode dans sa globalité, sans nécessairement se pencher sur la « boite noire » de son fonctionnement. Un certain nombre de ces travaux insiste ce que l'on nomme dans la littérature la « validité externe », c'est-à-dire la capacité des résultats des expérimentations

aléatoires à tenir dans d'autres contextes que celui dans lequel ils ont été obtenus. De plus, ces critiques, pour pertinentes et inspirantes qu'elles soient, se contentent, à de rares exceptions près (Bédécarrats, Guérin et Roubaud, 2015 ; Quentin et Guérin, 2013), d'aborder la question de la preuve sous un angle essentiellement théorique.

Dans cet article, nous pensons apporter une contribution originale à la littérature à travers la conjugaison de deux points : une approche « par le bas » (Hibou et Samuel, 2011a), qui vise à saisir la construction de la preuve dans les expérimentations aléatoires telles qu'elles se font et non telles qu'elles sont censées se faire, et une interrogation qui porte sur la qualité de la preuve produite, que l'on rencontre dans littérature sous le terme de « validité interne », c'est-à-dire la capacité d'un dispositif expérimental à mesurer précisément les effets d'un « traitement » (à distinguer des effets qui auraient une autre origine, comme le contexte par exemple). Nous nous penchons ainsi sur la validité *réelle* des expérimentations aléatoires. Que prouvent, *en pratique*, les expérimentations aléatoires ? Le passage de la théorie au terrain est-il à la hauteur des promesses de « hard evidence » que formulent ses promoteurs ? La validité interne est-elle si solide qu'il n'y paraît ?

Pour répondre à ces questions, nous avons mené ce que l'on pourrait qualifier d'enquête épistémologique, en cherchant à mettre au jour l'épistémologie des économistes pratiquant les expérimentations aléatoires[1], non seulement telle qu'elle pourrait se donner à voir à travers la lecture de leurs travaux, mais aussi, et surtout, en saisissant comment elles se traduisent sur le terrain. Pour commencer, nous avons mené une revue systématique de la littérature randomiste (Jatteau, 2013a), afin de connaître notre objet en profondeur. Le caractère extrêmement internationalisé du laboratoire, avec des expérimentations menées sur tous les continents et des chercheurs d'horizons variés (Jatteau, à paraître), dans des zones parfois reculées, complique les observations de terrain, sans compter la méfiance dont certains randomistes font montre. C'est pourquoi nous avons principalement utilisé comme matériaux des entretiens, menés avec différents acteurs de la randomisation, c'est-à-dire pas uniquement les chercheurs. En particulier, nous avons privilégié les assistants de recherche, dont le rôle est de faire l'interface entre les chercheurs, basés dans des universités occidentales, et le terrain, le plus souvent dans des pays pauvres et dans des zones rurales. Parce qu'ils y restent de longues périodes, généralement au minimum une année, ce sont des observateurs particulièrement attentifs du déroulement des expérimentations aléatoires. Leur témoignage sont ainsi précieux pour saisir comment, en pratique, la randomisation est appliquée.

Nous avons ainsi pu mettre en avant deux aspects qui questionnent directement la qualité de la preuve telle que produite par les expérimentations aléatoires. La production de chiffres, aboutissement de toute expérimentation aléatoire, apparaît largement décontextualisée, ce qui pose directement la question de ce qui est *réellement* prouvé (I). La réflexivité réduite dont font preuve les randomistes ne leur permet pas, à notre sens, de saisir pleinement le sens des données produites, et de la portée de la preuve qu'ils entendent fournir, notamment à cause de la dévalorisation des méthodes qualitatives (II).

[1] Dans la suite de l'article, nous les nommerons simplement « randomistes ».

I. De quoi parle la preuve ? Une production de chiffres décontextualisés

Le succès des expérimentations aléatoires reposent en partie sur la valeur sociale des chiffres (Porter, 1995), très élevée dans nos sociétés (Bruno et Didier, 2013). Elle n'est pourtant de sens qu'inscrite dans le contexte de leur production (Desrosières, 2008), pourtant négligés par les randomistes.

A. Les « réalités du terrain » souvent négligées

Le terrain occupe une place particulière dans les discours de tous ceux qui participent aux expérimentations aléatoires, que ce soit les chercheurs ou les assistants de recherche. Beaucoup mettent en effet en avant le côté « terrain » comme un progrès par rapport aux expériences de laboratoire ou à certains économistes du développement et des politiques publiques. Mais qu'entend-on ici par « terrain » ? Il n'est pas considéré comme potentiel pourvoyeur de connaissances à part entière, à même de contextualiser les résultats obtenus à l'aide de la randomisation, mais est vu comme purement utilitaire :

> « – Quand tu dis que tu pars sur le terrain avec eux, ça consiste en quoi ?
>
> – En gros c'est pour contrôler, vérifier qu'ils ont bien compris les questionnaires et tout, moi je rejoignais une équipe et je restais trois-quatre jours avec eux, et donc on était dans leur 4x4, et je faisais le trajet avec eux sur le terrain, je contrôlais, et le soir on débriefait, etc. » (Bettina, assistante de recherche)
>
> « – On the field, what are you doing?
>
> – It depends. So for example my last visit to [X] involved training our field staff on the health protocol [...] and some meetings with the ministry of health. In general, just to give some feedbacks to our staff members. » (Noémie, chercheuse)
>
> « [Telle chercheuse] elle est venue, elle est venue une fois avec moi [sur le terrain], pour nous aider à mettre en place le truc quoi. Mais c'est tout. Pas pour réfléchir au terrain, observer, voir comment ça se passe. Non. » (Kenza, assistante de recherche)

Aller sur le terrain, c'est donc essentiellement aller *contrôler* le déroulement de l'expérimentation, afin, précisément, que le terrain *ne perturbe pas* les données. Des réunions sont organisées avec les enquêteurs de terrain, qui sont chargés d'aller dans les villages distribuer le « traitement » et faire passer les questionnaires, ainsi qu'avec les assistants de recherche. Souvent, les chercheurs vont dans les villages où ont lieu les expérimentations, à titre d'exemple, pour avoir un aperçu. La démarche n'est cependant pas de s'intéresser au terrain en lui-même, mais de faire en sorte que l'expérimentation se passe bien d'un point de vue technique, presque logistique. Le terrain ici se limite donc le plus souvent aux bureaux locaux du laboratoire, agrémenté d'excursions dans les zones où le « traitement » est appliqué. Comme nous l'écrivions : « Dans l'acception des randomistes, le "terrain" correspond à un espace défini où l'objectif ultime n'est aucunement de faire de l'ethnographie, mais bien de recueillir des données auprès des individus qui participent à une expérimentation afin de produire des résultats chiffrés. Autrement dit, le terrain s'apparente à ce que l'on pourrait appeler un "terrain de jeu statistique". » (Jatteau, 2013b).

Cette vision purement technique du terrain occasionne des reproches de la part des assistants de recherche qui se trouvent, eux, sur place à l'année. Pour beaucoup, les chercheurs viennent trop peu. Tel assistant de recherche n'a jamais vu sur le terrain un chercheur qui dirige une expérimentation. Tel randomiste se rend quelques jours sur place, mais uniquement dans les locaux du laboratoire et non dans les villages.

Ainsi, les « réalités du terrain », c'est-à-dire les conditions de mise en œuvre d'une expérimentation, le contexte où elle s'inscrit, la manière de procéder, etc., semblent souvent négligées (Quentin et Guérin, 2013). Les assistants de recherche eux-mêmes relèvent des indices leur donnant à penser que les chercheurs qui dirigent les expérimentations sont bien souvent éloignés des contraintes qui se posent sur place. Leur présence en permanence dans les pays dans lesquels se déroulent les expérimentations leur rend facilement perceptible ce « décalage » :

> « Ca c'est un des trucs qui m'a le plus interpellé, c'était le rapport au terrain des économistes qui n'ont pas du tout l'habitude du terrain et qui eux se permettaient d'arriver là et de dire "ben voilà, nous, on est l'École d'Économie de Paris, on veut telles données et vous nous les donnez !". » (Kenza, assistante de recherche)

> « Souvent ce que les [assistants de recherche] reprochent, en tout cas moi c'est ce que je reprochais, c'est que les [chercheurs] ont pas forcément la vision des obligations du terrain. » (Adrien, assistant de recherche)

> « Y avait un côté, ils étaient vraiment pas très proches de leur terrain. Je pense ça c'est une vraie limite. Tu te dis y a un moment [un chercheur en vue du laboratoire] qui a jamais foutu les pieds là-dessus et qui répondait mais complètement à côté de la plaque à nos questions. » (Brigitte, assistante de recherche)

Cette relative méconnaissance des contraintes du terrain implique que des éléments importants de l'expérimentation ne fassent pas sens localement. Une large majorité des expérimentations aléatoires ont lieu dans des pays pauvres, et bien souvent dans des zones rurales. Le niveau d'éducation y est souvent faible, ce qui entraîne des difficultés de compréhension langagière à même de biaiser les résultats que l'on peut tirer d'une expérimentation. Que mesure-t-on vraiment quand les participants à une expérimentation peinent à comprendre les questions qu'on leur pose ?

> « Quand tu leur poses des questions sur leur chiffre d'affaires et qu'y a deux vaches, une chèvre, tu vois, tu te dis, c'est vraiment un truc qui a été dessiné dans des laboratoires américains et qui est pas en phase avec le terrain, c'est pas possible quoi. Y a des fois t'es un peu écœurée, tu te dis, on peut pas traduire ça en chiffres, t'en peux plus d'essayer de savoir dans quelle case ça rentre et d'unifier auprès de tes 35 enquêteurs la manière de poser la question. Est-ce que la chèvre est un capital fixe ou un stock ? T'en sais rien au final. » (Brigitte, assistante de recherche)

Cette question sur les chèvres comme capital fixe, loin d'être anecdotique, illustre cette distance au terrain. Là où dans un contexte académique la question conserve une certaine pertinence, elle apparaît totalement déplacée dans le cadre rural de ce pays africain. Cet exemple met également en avant la place, ou plutôt l'absence de place, de ceux que l'on pourrait appeler les « enquêtés » (par défaut, car ils ne sont généralement pas nommés), ou plus spécifiquement ici, les « randomisés ».

Il s'agit bien de mettre le terrain à distance, comme si les particularités qui ne manqueraient pas de surgir à l'étudier de trop près risquaient de fausser l'universalité de la méthode. Puisque

celle-ci est à même de s'appliquer en tout temps et en tout lieu, pourquoi donc s'intéresser au terrain ? Quel intérêt y aurait-il à contextualiser des résultats alors que la méthode noie toute différence dans le tirage au sort de deux groupes ? Sans doute parce qu'*en réalité*, il y des frottements, des ajustements, des arrangements. Normaux et systématiques dans chaque enquête, ils doivent être documentés et investigués, sous peine de ne pas saisir *réellement* « ce qui se passe », et au risque de ne plus faire des « randomized controlled trials » mais des « randomized out of controlled trials », comme le note un assistant de recherche (Jatteau, 2013b).

Dans les paragraphes qui précèdent, nous avons souligné le manque de considération pour le contexte et pour la réalité du « traitement ». Il nous faut à présent évoquer la conception épistémologique sous-jacente des randomistes, qui met le terrain à distance. De fait, si les expérimentations aléatoires sont bien des expérimentations « de terrain », tout l'enjeu est de savoir de quel terrain il s'agit. Les randomistes lui donnent un sens bien particulier, et comme ils le reconnaissent eux-mêmes, éloigné de celui des sociologues :

> « – Justement vous dites "ce qui est intéressant, c'est quand on va sur le terrain", c'est quelque chose que font pas tous les économistes.
>
> – Que fait personne, soyons sérieux.
>
> – Chez les économistes.
>
> – Chez les économistes. Evidemment, les sociologues en font, c'est normal, c'est leur métier. » (Stéphane, chercheur)
>
> « Parfois on va sur le terrain vraiment, voir [comment ça se passe]. Moi je l'ai fait à X. Mais on n'y est pas 24h/24, on peut pas dire que l'on a fait vraiment une observation participante comme un sociologue peut le faire. » (Octave, chercheur)
>
> « On discute avec eux, on fait cet exercice, qui évidemment vaut absolument pas ce que ferait un sociologue en allant sur le terrain, qui est plutôt quelque chose qui nous sert nous à avoir une idée d'ensemble de ce que c'est que ce dispositif, de l'ambiance, de comprendre un peu ce qui se passe. On se livre pas à un exercice approfondi de travail sur le terrain comme vous vous feriez. C'est pas notre métier et on cherche pas à le faire. » (Nathan, chercheur)
>
> « Le terrain en économie est complètement différent du terrain socio ou anthropo. Ce n'est pas pareil. » (Béatrice, assistante de recherche)

Il ne s'agit pas pour nous de prendre part à un débat théorique sur l'épistémologie de l'économie ou des sciences sociales en général, mais plutôt de rendre compte de comment les options épistémologiques des randomistes, consistant à tenir le terrain à distance en le « sous-traitant », sont à même de limiter l'intérêt des expérimentations aléatoires elles-mêmes car cela mine leur validité interne. En se privant des moyens de saisir le déroulement pratique d'une expérimentation et d'agréger des éléments de contexte, on manque d'éléments de compréhension.

Ainsi, quand Adrien, un assistant de recherche, dit qu'il a « rarement vu un truc plus de terrain qu'une RCT [« randomized controlled trial », expérimentation aléatoire en anglais] en fait », c'est révélateur de ce qu'est un terrain pour bon nombre d'acteurs des expérimentations aléatoires. Le sens est technique et géographique : le terrain, c'est la terre, celle où l'on met les pieds. Faire du terrain, c'est poser le pied là où se déroulent les expérimentations, sans véritable réflexivité, et donc sans nécessairement saisir ce qui s'y passe. Or cette décontextualisation de la production chiffrée dans le cadre des expérimentations aléatoires pose des risques heuristiques. Que peut-on

vraiment dire d'un « traitement », si on méconnaît le contexte de l'évaluation et si on n'a pas une connaissance fine de ce en quoi il consiste *réellement* ? Comment tirer des enseignements d'un terrain qu'on refuse de considérer ?

Cette dévalorisation des « réalités du terrain » invite directement à nuancer la validité interne. Si l'on ne sait pas vraiment « ce qu'il se passe », comment peut-on simplement l'évaluer ? Tout se passe comme si la séduction qu'opérait le modèle théorique de la randomisation aveuglait les économistes qui y ont aujourd'hui recours et ne leur permettait pas de voir ce qu'impliquait le passage à la pratique.

B. Des « traitements » mal connus

Cette option épistémologique – qui consiste à « laisser » le terrain, en le déléguant aux assistants de recherche et en n'en faisant pas un point nodal de la démarche – a donc des implications très concrètes. Nous souhaitons insister sur l'une d'elles en particulier : le « traitement ». Aussi clair puisse-t-il être en théorie, c'est-à-dire au moment de sa conception, sa traduction sur le terrain peut s'en éloigner, car il faut alors composer avec les « réalités du terrain », faites d'obstacles et de négociations. Paradoxalement, l'une des difficultés de ce genre d'expérimentations est précisément de savoir en quoi a réellement consisté le « traitement » (Bamberger et White, 2007). Plusieurs chercheurs nous ont ainsi confié ne pas savoir *réellement* en quoi consistait le traitement, ou alors comment il se déroulait *en pratique* ce qui soulève la question de la validité interne de la méthode.

Beaucoup d'assistants notent qu'ils connaissent pour la plupart bien mieux le « traitement » que les chercheurs qui dirigent le projet et, au-delà, le terrain. Ils sont d'ailleurs particulièrement critiques sur cette méconnaissance du programme qui est réellement testé (et non de celui qui était prévu).

Cette connaissance du « traitement » est fondamentale. En effet, il ne s'agit pas d'assister à sa mise en œuvre afin de donner de la « chair », pour reprendre le terme d'un randomiste, et pour peaufiner la description que l'on fera dans les publications futures, mais bien de savoir ce que l'on l'évalue en pratique, et non plus en théorie :

> « – Mais du coup en fait […] dans ces réunions [qui constituent le traitement], ni toi, ni les chercheurs n'y assistaient en fait.
>
> – Non, personne n'y assistait.
>
> – Donc on sait pas trop [ce qu'a été précisément le traitement] ?
>
> – On sait pas. On sait pas. » (Kenza, assistante de recherche) »

> « – Et juste dernière question parce que ça c'est difficile de trouver des documents dessus, juste en dernier, tu peux me dire un peu précisément en quoi consistait le traitement, qu'est-ce que l'on leur apprenait, comment ça se passait dans la pratique, etc. ?
>
> – […] Je sais pas y avait un volet coaching, un volet cours de comptabilité, y avait des forums et ils les mettaient en relation. Ca aussi putain... Ca je sais pas dans quelle mesure les chercheurs ont eu un impact dans la manière dont ce truc-là s'est fait. C'était genre ils font, ils organisent leur trajet de leur petit village complètement pourri, enfin complètement isolé disons très pauvre etc., jusqu'à [une grande ville] où ils les mettent dans un super hôtel pendant 3 jours. Tu vois... Genre, il est où le traitement là-dedans ? Les types le seul truc dont ils se souviennent c'est "c'était super, l'hôtel il était incroyable". » (Brigitte, assistante de recherche)

Ce dernier exemple illustre bien le problème que sous-entend le manque de caractérisation du « traitement » : qu'évalue-t-on ? De quoi mesure-t-on les effets ?

> « – Tu crois pas aux résultats ?
>
> – Non. Parce que tu te dis qu'y a pas eu de... Bah, tu sais même pas que t'évalues, parce que tu sais pas le contenu [du traitement]. » (Kenza, assistante de recherche)

Les zones où se déroulent les évaluations sont rarement totalement vierges de toute politique développementale et les rapports avec les agents, passés et présents, doivent être pris en compte pour apprécier les effets d'un nouveau programme. Il faut « prendre en compte "l'histoire locale" des contacts avec l'interventionnisme politico-économique » (Olivier de Sardan, 1995), car cela renvoie directement à la caractérisation du traitement. Il s'agit en effet de savoir précisément d'où provient l'effet que l'on mesure : est-il uniquement dû au « traitement » ? Que contient le « traitement » lui-même ? Qu'est-ce que le « traitement » ? Il est possible de considérer le « traitement » comme embarquant cette histoire locale. De fait, ce n'est pas seulement « la distribution gratuite de moustiquaires » ou « l'incitation à se faire vacciner » que l'on teste. Ce sont ces programmes dans leur contexte, qu'il convient donc de préciser, car ce dernier fait partie intégrante du « traitement ».

De fait, il apparaît que dans un certain nombre de cas, la connaissance du « traitement » ne paraît pas suffisante pour documenter de manière suffisamment précise le programme évalué.

II. De quoi fait-on preuve ? Une réflexivité réduite

L'absence de prise en compte du contexte de production des expérimentations aléatoires renvoie à la question de la réflexivité. Le recul nécessaire à toute enquête, qu'elle soit de nature qualitative ou quantitative, semble être trop peu présent et met directement en cause la qualité de la preuve obtenue.

A. Une incapacité à saisir le passage de la théorie à la pratique

La réflexivité, au sens de retour sur ses pratiques, permet de mieux saisir comment celles-ci, par les nécessaires ajustements qu'elles impliquent, influencent le protocole que l'on s'est fixé et, partant, les conclusions que l'on peut en déduire.

Confrontés au terrain, les chercheurs adaptent leur méthodologie, divers « arrangements » sont trouvés. Pour les expérimentations aléatoires, cela peut porter sur la procédure de tirage au sort (en la rendant publique par exemple) pour essayer de dépasser les réticences des enquêtés, sur le choix de tel indicateur, de telle catégorie de population pour faire l'expérience, etc. Ces choix ne sont pas neutres (Gabas, Ribier et Vernières, 2013), il faut donc être attentif aux « effets de catégorisation » (Hibou et Samuel, 2011b), c'est-à-dire à l'existence de ce qui apparaît comme des impacts, mais qui pourraient se limiter aux catégories auxquelles on a recours pour les mesurer. Les indicateurs mobilisables peuvent être nombreux pour un même programme et, vu le seuil de significativité généralement considéré (5 %), les chances sont élevées pour trouver, parmi eux, des résultats « significatifs ».

La réflexivité doit ici également être pensée vis-à-vis des pratiques qui sont propres au(x) terrain(s). Ceux-ci sont souvent éloignés des lieux de vie et de travail des chercheurs, et ces

derniers entretiennent, comme on l'a vu, une certaine distance à leur égard. A ce titre, le travail de contextualisation apparaît incontournable (Jerven, 2013), et ce d'autant que les terrains peuvent être lointains, comme en Afrique, où « la difficulté est […] exacerbée dans des contextes éloignés des "cadres sociaux" dans lesquels les concepts ont été élaborés » (Hibou et Samuel, 2011b).

Le travail de contextualisation est souvent réalisé par les randomistes dans le début de leur article, mais il s'agit d'une contextualisation particulière, qui se borne le plus souvent à une description de l'environnement institutionnel autour de l'expérimentation. On présente le fonctionnement du système scolaire du pays, on donne quelques statistiques sur les agriculteurs de la région, etc. Mais la réflexion ne porte pas véritablement sur le terrain lui-même, en lui-même et pour lui-même, notamment concernant les « frottements » que toute enquête entraîne.

La réflexivité concernant les acteurs de l'enquête, en premier lieu les enquêteurs de terrain, est traitée généralement de façon partielle car quantitative. L' « effet enquêteur », bien connu des assistants de recherche comme des chercheurs, et qui veut qu'un enquêteur en particulier, par la manière qu'il a de poser les questions lors du recueil de données puis ensuite de les coder a des chances de biaiser les réponses, n'est appréhendé que par une variable supplémentaire dans la régression linéaire. Il n'est pas question d'aller voir de plus près comment se déroulent les interactions entre les enquêteurs et les enquêtés. Celles-ci apparaissent captables par le modèle de régression sur lequel s'appuient presque tous les articles relatant des expérimentations aléatoires. Or les travaux sur cette question montrent à quel point les enquêteurs peuvent être amenés à « s'arranger », en adaptant les questions, en ne lisant pas toutes les réponses, voire parfois en y répondant eux-mêmes, en « bidonnant » (Caveng, 2012). Il ne s'agit pas seulement d'un « effet enquêteur » mais d'un « effet des enquêteurs ». Comme l'écrit Rémy Caveng à propos des instituts de sondage, « les écarts aux prescriptions sont donc essentiellement perçus comme des risques d'altération de la qualité des données et très rarement comme des contributions essentielles en termes de quantité et de qualité de l'information recueillie » (Caveng, 2012).

C'est comme si loin d'être un allié dans la constitution de connaissances, cette « mise en pratique » agissait comme un ennemi. Les expérimentations aléatoires sont bien des expériences « de terrain », mais les randomistes agissent en bonne partie en faisant « comme si » elles étaient des expériences de laboratoire.

La formule désormais classique d'Alain Desrosières, selon laquelle « quantifier, c'est convenir puis mesurer » (Desrosières, 2008), n'est guère reprise à leur compte par les randomistes. Les choix ne sont pas systématiquement explicités. Aucune mise au jour de ce qui est incorporé, routinisé, codifié, n'est opérée, ce qui entraîne une naturalisation de ce que l'on mesure. Les chiffres produits, *même* par randomisation, c'est-à-dire *même* avec un degré de scientificité prétendument supérieur à celui des autres méthodes, ne sauraient être un pur reflet de la réalité (Desrosières, 2000), d'où la nécessité de reconstituer leur genèse.

Pour donner un exemple de ce qu'une prise en compte très relative du terrain peut donner, on peut évoquer l'épicentre de la vague des expérimentations aléatoires dans le monde du développement, qui se situe à Busia, au Kenya. C'est là que les premières ont eu lieu à la fin des années 1990, auxquelles ont succédé des dizaines d'autres. A l'époque où nous étions assistant de recherche, le bureau du J-PAL chargé de gérer sur place les expérimentations aléatoires était un des plus gros employeurs locaux, et toute la ville connaissait son existence. Les expériences menées aux alentours étaient nombreuses, ce qui n'est pas sans conséquence :

> « Et ça c'est une grosse critique dans les pays en développement et surtout sur des zones au Kenya qui sont "overresearched" notamment la zone où on était en province de l'Ouest, c'est que les répondants ont été par 36 000 manières enquêtés pour différents programmes que ça soit [nom d'un projet W], que ça soit [nom d'un projet X], que ça soit [nom d'un projet Y], que ça soit [nom d'un projet Z], donc en quoi tu peux dire que vraiment tu évalues vraiment un traitement à proprement parler quand ces personnes-là sont habituées à être enquêtées, sont habituées à recevoir des cadeaux par rapport aux enquêtes et tout ça. Ça c'est... Je te parle d'un truc qui est propre au Kenya et à l'expérimentation que l'on a menée, mais c'est quelque chose qui m'a choquée vraiment, c'était l'aspect "overresearched" quoi. » (Béatrice, assistante de recherche)

Ce problème n'est pas généralisable à toutes les expérimentations, certaines étant isolées et intervenant sur des terrains « vierges ». Mais dans de nombreux endroits, par souci d'économie, un certain nombre de projets sont lancés à partir d'une même base. Dès lors, si les populations alentour sont multi-enquêtées, se pose la question de la validité interne des expérimentations en cours. Qu'évalue-t-on quand on évalue un programme auprès d'individus qui ont déjà été sensibilisés à d'autres programmes auparavant, qui ont déjà été enquêtés, etc. ? Ce n'est pas problématique en soi, pour peu que cela soit pris en compte dans l'analyse, ce qui ne semble guère être le cas. Pour y parer, on en vient à envisager des solutions *ad hoc*, comme masquer son appartenance au J-PAL sur le terrain, afin que les enquêtés ne réagissent en fonction des précédentes expérimentations desquelles ils ont fait partie.

Il nous faut souligner que ce n'est absolument pas l'existence même d'arrangements, incontournables dans la recherche, particulièrement dans la recherche appliquée comme le sont les expérimentations aléatoires, qui constitue une éventuelle limite de ces dernières. Mais c'est la faible prise en compte de ces arrangements par les randomistes qui pose question (Bédécarrats, Guérin et Roubaud, 2015).

Il ne faut en effet pas « taire » ces bricolages, qui, une fois mis au jour et analysés, sont à même d'accroître la scientificité des enquêtes (Quentin et Guérin, 2013). L'enjeu ici n'est pas de se conformer *par principe* à une *doxa* des sciences sociales, qui voudraient rendre nodale la réflexivité – à raison d'ailleurs –, mais bien de se rendre compte que son absence entache la « validité interne », dont la randomisation serait la championne. Une fois appliquée, cette méthode ne souffre pas de moins d'arrangements que les autres. Dès lors, quand bien même le tirage au sort des groupes permettrait d'éviter les biais de sélection, si l'on ne sait pas *réellement* ce que l'on évalue, si le « traitement » effectivement donné n'est pas précisément connu, si la passation des questionnaires n'est pas bien documentée, si la construction des indicateurs n'est pas suffisamment questionnée, alors on en sait bien peu sur ce que l'on mesure. Autrement dit, s'il y a bien un effet, il est en réalité bien difficile de savoir *à quoi* l'attribuer.

B. Des bricolages de diverses natures

Les « bricolages » peuvent être de plusieurs types, certains étant communs aux enquêtes (comme les questionnaires), d'autres spécifiques aux approches quantitatives (comme les indicateurs) et d'autres encore aux expérimentations aléatoires.

C'est le cas du tirage au sort. On pourrait penser que ce point soit le plus standardisé, celui dont l'application pose le moins de problèmes. En réalité, il existe plusieurs façons de réaliser ce tirage au sort, que ce soit sur un plan technique (par un algorithme, par un tirage dans une urne à

la vue de tous, par un lancer de dés…) ou méthodologique (l'enjeu étant alors de savoir si l'on randomise au niveau individuel ou groupal).

En pratique, la randomisation n'est pas toujours aussi rigoureuse qu'en théorie. Angus Deaton remarque que dans la célèbre expérimentation des vermifuges (Miguel et Kremer, 2004), Edward Miguel et Michael Kremer n'ont pas procédé à un tirage au sort *stricto sensu* (Deaton, 2010), à cause de résistances sur place. Ils ont dû se contenter de ce qui pourrait être qualifiée de « quasi-randomisation », puisqu'ils ont assigné les écoles en fonction de l'alphabet (avec comme modèle : école A dans le groupe 1, école B dans le groupe 2, école C dans le groupe 1, etc.). Le canon de la randomisation n'est donc pas strictement respecté ici (Glennerster et Takavarasha, 2013). Aurélie Quentin et Isabelle Guérin signalent quant à elles, à propos d'une expérimentation aléatoire au Cambodge, divers bricolages et manquements au tirage au sort (Quentin et Guérin, 2013), tout comme Bernard Gomel et Evelyne Serverin à propos de celle sur le placement des chômeurs (Gomel et Serverin, 2013). Pour l'expérimentation du Revenu de Solidarité Active, le tirage au sort n'a tout simplement pas été autorisé, pour des raisons aussi bien éthiques que pratiques. Miriam Bruhn et David McKenzie étudient la façon dont la randomisation a été effectuée dans un large échantillon d'articles relatant des expérimentations aléatoires (Bruhn et McKenzie, 2009). Ils remarquent tout d'abord que peu de papiers explicitent la manière dont le tirage au sort a été opéré. De plus, quand ils interrogent des chercheurs sur la meilleure façon de randomiser concernant quelques cas typiques, ceux-ci fournissent des réponses divergentes, ce qui montre que la méthode, en son cœur, n'est pas aussi stabilisée que l'on pourrait le penser. D'où l'importance d'exercer une certaine réflexivité en la matière. Le tirage au sort lui-même est un construit social que l'on ne saurait prendre pour donné et qu'on ne peut donc réduire à une simple ligne de commande dans un logiciel de traitement de données.

Les propos que nous avons recueillis auprès des assistants de recherche nous indiquent d'autres difficultés qui conduisent à certains bricolages. Lorsque le niveau de langue des questionnaires apparaît trop compliqué, soit pour être bien saisi par les enquêteurs de terrain qui les passent, soit parce que ces derniers anticipent une incompréhension de la part des enquêtés, alors des questions peuvent être reformulées, suivant l'appréciation des enquêteurs eux-mêmes. A propos d'une expérimentation aléatoire qu'elles ont suivie, Aurélie Quentin et Isabelle Guérin notent que la longueur et la complexité du questionnaire était critiquée par les acteurs présents sur place de longue date (Quentin et Guérin, 2013).

La « déstandardisation » que peuvent opérer les enquêteurs n'est pas antinomique des intérêts des chercheurs, dans la mesure où bien souvent, c'est elle qui rend possible l'enquête et un taux de réponses jugé acceptable. Certains questionnaires durent normalement plusieurs heures s'ils sont menés jusqu'au bout et, surtout, dans les règles de passation prévues (lecture en entier de chaque question et de chaque réponse possible, par exemple). De fait, dans ces différentes « adaptations », les enquêteurs de terrain « élaborent et appliquent des conventions qui ne sont ni celles des concepteurs ni celles des analystes » (Caveng, 2012). L'existence de ce différentiel de conventions touche directement à la validité interne des expérimentations aléatoires. Méconnaître le mode de passation des questionnaires, la formulation de certaines questions, leur interprétation par ceux qui les codent, c'est aussi ne pas saisir de manière suffisamment précise « ce que les données veulent dire » et, par-là, quel est l'impact du programme testé.

La question des indicateurs utilisés pour mesurer la présence d'effets d'un programme est également cruciale. En macroéconomie, cela peut être spectaculaire, avec par exemple le doublement du PIB par habitant du Ghana… en une nuit, suite à des changements de normes comptables (Jerven, 2013) ! Or les agrégats macroéconomiques donnent une image du pays et peuvent avoir des effets tout à fait réels, notamment en termes d'accès à des bailleurs internationaux par exemple, dont l'aide est de plus en plus conditionnelle. En matière de développement, le choix des indicateurs est donc particulièrement primordial, puisqu'en fonction de ce que l'on regarde, on peut mesurer une présence ou au contraire une absence de résultats. Nombreuses sont les expérimentations aléatoires où les *outcomes* peuvent être multiples et où une discussion s'impose pour les choisir et débattre de leur pertinence mais aussi de leurs limites. Les indicateurs embarquent avec eux un ensemble de conventions, pas toujours explicitées, et « s'inscrivent dans des courants de pensée » (Gabas, Ribier et Vernières, 2013). Comme l'écrit Catherine Paradeise, « les indicateurs sont porteurs de visions de la réalité que l'on confond, en régime ordinaire, avec "la réalité" » (Paradeise, 2012). Ce « portage » est normal et habituel, mais une démarche réflexive s'impose pour en rendre compte. « L'indicateur traduit, et à ce titre il trahit », pour reprendre la formule de Catherine Paradeise. C'est cette traduction qui nous intéresse ici et qui est trop passée sous silence par les randomistes. La trahison ne s'opère que si les clés de la traduction ne sont pas données et si ces dernières sont réifiées de telle sorte que l'on ne puisse plus en établir la genèse. Il en va de la confiance que l'on peut accorder aux chiffres ainsi produits.

C. Des approches qualitatives déconsidérées

Les sociologues et les anthropologues sont familiers avec les problèmes que nous venons de soulever. Ils ont recours à des méthodologies qui permettent de prendre en compte les spécificités du terrain et d'opérer une réflexivité sur leurs pratiques. Elles sont souvent de nature qualitative (Olivier de Sardan, 2008), ce qui limite leur usage par les randomistes.

En effet, la hiérarchie des preuves postulée par les randomistes est double. Celles provenant des méthodes quantitatives sont supérieures à celles de nature qualitative, et parmi les quantitatives, la randomisation occupe la première place. Il ressort clairement, aussi bien de leurs discours que de leurs travaux de recherche, une dévalorisation du qualitatif, qui va même jusqu'à une déqualification. À leurs yeux, le « chiffre randomisé » n'est pas seulement le plus haut niveau de preuves, il est le seul *pleinement* scientifique alors que le qualitatif ne l'est, lui, *pas du tout*. A la mise à distance du terrain s'ajoute celle des approches qualitatives, qui ne sont pas reconnues comme appartenant en tant que telles à la discipline économique :

> « – Quand Michael Kremer [un économiste très reconnu du J-PAL] au repas, quand il m'a demandé sur quoi j'avais fait ma thèse et tout, je lui ai dit j'ai fait une thèse en économie tout ça et j'ai raconté ce que j'ai fait, il m'a dit "c'était super donc en fait t'as fait une thèse en anthropologie" .
>
> – Donc si y a du quali, c'est pas de l'économie en fait ?
>
> – En quelque sorte. Mais lui il est d'accord avec ça, il est d'accord avec tous les gens qui font du quali, c'est juste que voilà c'est pas la même discipline. Je crois pas qu'il critique, c'est juste que c'est pas la même discipline en fait. » (Adrien, assistant de recherche)

Cette délégitimation des autres modes de savoir se double logiquement d'une ignorance des autres littératures (Barrett et Carter, 2010). Il est très rare de trouver dans la littérature randomiste des références provenant d'autres champs disciplinaires, à moins que ceux-ci ne se réfèrent à une administration de la preuve sur un mode quantitatif et plus particulièrement randomisé, comme une partie de la science politique.

Cette déconsidération se retrouve dans la pratique des expérimentations aléatoires. Les sommes puisées dans les budgets pour « faire du qualitatif » sont marginales. Peu de temps y est consacré. Un assistant de recherche que nous avons rencontré raconte ainsi qu'il était censé réaliser, après le questionnaire, une « note de terrain » élaborée notamment à partir d'une « discussion » avec l'enquêté. Mais le questionnaire était déjà si long qu'il ne lui restait le plus souvent pas suffisamment de temps, ce qui illustre bien la place dévolue au qualitatif. Ce dernier est presqu'entièrement « sous-traité » aux assistants de recherche – à eux de « faire le travail de terrain ». Il n'est pas réalisé suivant les canons méthodologiques propres aux sciences sociales comme la sociologie ou l'anthropologie et il apparaît comme une intrusion dans le protocole bien huilé de l'expérimentation aléatoire. Quand on fait appel à des collaborateurs pour intervenir *en marge* de l'expérimentation dotés de connaissances spécifiques dans les méthodes qualitatives (ce qui demeure très rare), ils arrivent sur le terrain « comme un cheveu sur la soupe », pour reprendre l'expression d'une assistante de recherche.

Cette dévalorisation des méthodes qualitatives traduit une épistémologie particulière de l'économie, qui serait (devenue) une science expérimentale (Cahuc et Zylberberg, 2016). Comme nous l'a expliqué en entretien un chercheur : « […] Faire des entretiens, c'est le signe que l'on est moins dans la science que dans l'art. » Une rupture assez nette est faite entre l'économie et la sociologie (à laquelle on rattache systématiquement le qualitatif, ce dernier étant d'une certaine manière « orthogonale » à l'économie) :

> « – Donc là y a une vraie complémentarité ?
>
> – Ah bah complètement.
>
> – Vous mesurez l'impact et les sociologues...
>
> – Expliquent.
>
> – C'est intéressant parce que quand on regarde la littérature [des expérimentations aléatoires] surtout en développement, il est assez peu fait appel à des sociologues et à des anthropologues.
>
> – Souvent aussi vous savez on ne parle pas non plus complètement la même langue. Et c'était assez amusant parce que l'on avait fait une réunion de travail et y avait [des économistes] d'un côté, et [des sociologues de l'autre]. Et on se mettait d'accord sur un papier, parce que l'on voulait faire un papier ensemble pour essayer de le publier. Et on se disait "voilà partie 1 machin, partie 2 bidule, partie 3, etc." Et on s'en rend pas compte, mais on a peut-être même un esprit formaté, les uns et les autres, ils ne comprenaient pas comment on voulait l'articulation entre leur rapport et le nôtre. Vraiment, ils comprenaient pas, c'était un dialogue de sourds. […] C'est vrai que nous on est très simplification de la réalité, des outils assez lourds, très maths, on aime bien. Un économiste préfère une équation plutôt qu'un paragraphe rédigé. Alors que les sociologues sont très rédaction, très littéraire. » (Nadège, chercheuse)

Cette option épistémologique n'est pas propre aux randomistes et traversent une large part de la discipline contemporaine, quoique certains chercheurs en vogue ont tendance à la remettre en cause, comme Thomas Piketty. Le qualitatif apparaît ainsi peu défini et se situerait « de l'autre

côté », c'est-à-dire pas du côté de l'économie, ce qui fait que les randomistes n'en ont pas toujours une idée claire. Comme le dit une assistante de recherche : « Ils ont un groupe de recherche, ils sont gérontologues, sociologues, tout ça ». Le « tout ça » indique un certain flou autour de tous les « – ogues ». Notons que ce rejet du qualitatif est lui-même conforté par le système des publications en cours dans les grandes revues américaines (Jatteau, 2013b), où en ne lui accordant pas de place, on réduit singulièrement ses chances de publier. Il ne faut pas perdre de vue que les randomistes s'adressent aussi (et surtout ?) aux éditeurs de revues quand ils mènent une expérimentation. Si ceux-ci modifiaient leurs choix éditoriaux pour favoriser le qualitatif, il est permis de penser que les randomistes adapteraient leur pratique.

Conclusion. Une validité interne à nuancer fortement

Les promoteurs des expérimentations aléatoires ces dernières années formulaient avec une promesse : celle d'avoir trouvé un outil à même de fournir des preuves de qualité supérieure. La randomisation permettrait de mesurer des impacts avec un degré de certitude qu'aucune autre méthode n'atteint. Elle serait à même de « révolutionner les politiques publiques du XXI^{e} siècle » (Duflo et Kremer, 2008). Il est vrai que d'un pur point de vue méthodologique, l'utilisation du tirage au sort à des fins évaluatives a de quoi séduire et apporte un élément décisif en comparaison des méthodes usuelles dans l'évaluation quantitative d'impact, comme les doubles différences ou la régression par discontinuité (Jatteau, 2013a).

Malgré certaines critiques qui se sont faites jour, dont bon nombre reprennent celles déjà pointées dans les années 1970 aux Etats-Unis (Jatteau, 2016), force est de constater que les expérimentations aléatoires ont connu un succès croissant depuis le début des années 2000. Elles font désormais partie de la boite à outils classique de l'évaluateur. De plus, elles sont souvent considérées comme la meilleure méthode pour fournir des preuves scientifiques d'évaluation d'impact, aussi bien en économie du développement qu'en évaluation des politiques publiques, comme en témoigne par exemple le *Guide méthodologique pour l'évaluation des expérimentations sociales* du Fonds d'expérimentation pour la jeunesse (Conseil scientifique du FEJ, 2009).

Au-delà des précautions d'usage que permettent de formuler la sociologie de la quantification (Desrosières, 2008), l'approche originale que nous avons développée dans cet article – une approche « par le bas » de la preuve par le chiffre tel que produit par les expérimentations aléatoires – entend remettre en cause la supériorité de la randomisation en matière de validité interne. En étudiant attentivement comment ces chiffres sont produits, nous avons montré que leur décontextualisation et la relative absence de réflexivité des chercheurs nuançaient fortement la qualité de la mesure d'impact. La mise à distance du terrain opérée, souvent implicitement, limite sa prise en compte, ce qui a des conséquences directes sur les enseignements que l'on peut en tirer. Les « traitements », non pas « théoriques » mais tels qu'ils ont réellement eu lieu, eux-mêmes peuvent être mal connus. L'étude du passage de la théorie de la randomisation à la pratique expérimentale, avec ce qu'elle contient comme ajustements et comme bricolages, apparaît comme une faiblesse de la randomisation telle qu'elle est pratiquée aujourd'hui.

Ces manquements laissent donc penser que les chiffres produits par les expérimentations ne relèvent pas de preuves aussi solides que certains veulent bien le croire. Notre propos n'est pas de contester l'utilité même de cette méthode, mais d'en questionner la supériorité en matière de

production de preuves d'efficacité. La faiblesse de la prise en compte du contexte alliée à une réflexivité réduite limite donc la validité interne et, au-delà, l'intérêt même que l'on peut porter à cette méthode dans ses conditions actuelles d'utilisation.

Les questionnements dont il a été question dans cet article renvoient en creux à ceux plus larges portant sur le pluralisme méthodologique et disciplinaire. La mise à distance du terrain et la méfiance pour les approches qualitatives renvoient à une épistémologie particulière de l'économie et plus généralement des sciences sociales. Le refus inconscient de la réflexivité propre aux sciences sociales comme la dévalorisation des méthodes qualitatives soulignent l'absence d'ouverture disciplinaire et méthodologique. En la matière, les randomistes sont à l'image d'une partie de la science économique contemporaine, qui ne voit de scientifique que le quantitatif. Au moins concernant cette méthode, nous espérons avoir montré toute la fertilité qu'il peut y avoir à (ré)introduire un pluralisme méthodologique et disciplinaire.

La question de la possibilité d'expérimentations aléatoires que l'on pourrait qualifier de « pluralistes » reste ouverte. L'étude sociologique des randomises (Jatteau, à paraître) comme du fonctionnement du champ académique, notamment de la science économique (Fourcade, Ollion et Algan, 2015), laisse peu d'espoir en la matière. Cependant, en montrant les faiblesses qu'implique ce cloisonnement disciplinaire et méthodologique, nous pensons qu'une évolution de la manière de pratiquer les expérimentations aléatoires est souhaitable, à défaut d'être possible à l'heure actuelle.

Bibliographie

BAMBERGER M., WHITE H., 2007, « Using Strong Evaluation Designs in Developing Countries: Experience and Challenges », *Journal of MultiDisciplinary Evaluation*, *4*, 8, p. 58‑73.

BANERJEE, A.V. (dir.), 2007, *Making Aid Work*, Cambridge, The MIT Press.

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

BARRETT C.B., CARTER M.R., 2010, « The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections », *Applied Economic Perspectives and Policy*, *32*, 4, p. 515‑548.

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*, 27, p. 107‑122.

BÉDÉCARRATS F., GUÉRIN I., ROUBAUD F., 2015, « L'étalon-or des évaluations randomisées : du discours de la méthode à l'économie politique », Document de travail, DT 2015-01, Document de travail, DIAL.

BRUHN M., MCKENZIE D., 2009, « In Pursuit of Balance: Randomization in Practice in Development Field Experiments », *American Economic Journal: Applied Economics*, *1*, 4, p. 200‑232.

BRUNO I., DIDIER E., 2013, *Benchmarking. L'Etat sous pression statistique*, Paris, Zones, 209 p.

CAHUC P., ZYLBERBERG A., 2016, *Le Négationnisme économique et comment s'en débarrasser*, Flammarion.

CAVENG R., 2012, « La production des enquêtes quantitatives », *Revue d'anthropologie des connaissances*, *6*, 1, p. 65‑88.

CONSEIL SCIENTIFIQUE DU FEJ, 2009, « Guide méthodologique pour l'évaluation des expérimentations sociales », Fonds d'Expérimentation pour la Jeunesse.

DEATON A., 2010, « Instruments, randomization, and learning about development », *Journal of Economic Literature*, *48*, 2, p. 424‑455.

DEATON A., CARTWRIGHT N., 2016, « Understanding and Misunderstanding Randomized Controlled Trials », Working Paper, 22595, Working Paper, National Bureau of Economic Research.

DESROSIÈRES A., 2000, « L'histoire de la statistique comme genre : style d'écriture et usages sociaux », *Genèses*, *2*, 39, p. 121‑137.

DESROSIÈRES A., 2008, *L'argument statistique I. Pour une sociologie historique de la quantification*, Paris, Mines ParisTech-les Presses, 329 p.

DUFLO E., KREMER M., 2008, « Use of Randomization in the Evaluation of Development Effectiveness », dans EASTERLY W. (dir.), *Reinventing Foreign Aid*, Cambridge, MIT Press, p. 93‑120.

FOURCADE M., OLLION E., ALGAN Y., 2015, « The Superiority of Economists », *Journal of Economic Perspectives*, *29*, 1, p. 89‑114.

GABAS J.-J., RIBIER V., VERNIÈRES M., 2013, « Introduction. La mesure du développement : comment science et politique se conjuguent », *Revue Tiers Monde*, *1*, 213, p. 7‑22.

GLENNERSTER R., TAKAVARASHA K., 2013, *Running Randomized Evaluations: A Practical Guide*, Princeton, Princeton University Press.

GLEWWE P., KREMER M., MOULIN S., 2009, « Many children left behind? Textbooks and test scores in Kenya », *American Economic Journal: Applied Economics*, *1*, 1, p. 112‑135.

GLEWWE P., KREMER M., MOULIN S., ZITZEWITZ E., 2004, « Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya », *Journal of Development Economics*, *74*, 1, p. 251‑268.

GOMEL B., SERVERIN É., 2013, « L'expérimentation sociale aléatoire en France en trois questions », *Travail et emploi*, *3*, 135, p. 57‑71.

HIBOU, B., SAMUEL, B. (dirs.), 2011a, *La macroéconomie par le bas*, Paris, Karthala (Politique Africaine), 210 p.

HIBOU B., SAMUEL B., 2011b, « Macroéconomie et politique en Afrique », *Politique africaine*, 124, p. 5‑28.

JATTEAU A., à paraître, « Comment expliquer le succès de la méthode des expérimentations aléatoires ? Une sociographie du J-PAL », *SociologieS*.

JATTEAU A., 2013a, *Les expérimentations aléatoires en économie*, Paris, La Découverte (Repères).

JATTEAU A., 2013b, « Expérimenter le développement ? Des économistes et leur terrain », *Genèses*, *4*, 93, p. 8‑28.

JATTEAU A., 2016, *Faire preuve par le chiffre ? Le cas des expérimentations aléatoires en économie*, Thèse, ENS Paris Saclay.

JERVEN M., 2013, *Poor Numbers*, Londres, Cornell university press, 187 p.

LABROUSSE A., 2010, « Nouvelle économie du développement et essais cliniques randomisés : une mise en perspective d'un outil de preuve et de gouvernement », *Revue de la régulation*, 7, p. 2‑32.

LAURENT C., BAUDRY J., BERRIET-SOLLIEC M., KIRSCH M., PERRAUD D., TINEL B., TROUVÉ A., ALLSOPP N., BONNAFOUS P., BUREL F., CARNEIRO M.J., GIRAUD C., LABARTHE P., MATOSE F., RICROCH A., 2009, « Pourquoi s'intéresser à la notion d' "evidence-based policy" ? », *Tiers Monde*, 200, p. 853‑873.

MIGUEL E., KREMER M., 2004, « Worms: identifying impacts on education and health in the presence of treatment externalities », *Econometrica*, *72*, 1, p. 159‑217.

MONNIER E., 1992, *Evaluations de l'action des pouvoirs publics*, Paris, Economica.

OLIVIER DE SARDAN J.-P., 1995, *Anthropologie et développement. Essai en socio-anthropologie du changement social*, Paris, Karthala.

OLIVIER DE SARDAN J.-P., 2008, *La rigueur du qualitatif. Les contraintes empiriques de l'interprétation socio-anthropologique*, Louvain-la-Neuve, Academia Bruylant.

PARADEISE C., 2012, « Le sens de la mesure. La gestion par les indicateurs est-elle gage d'efficacité ? », *Revue d'économie du développement*, *26*, 4, p. 67‑94.

PORTER T.M., 1995, *Trust in numbers: the pursuit of objectivity in science and public life*, Princeton, Princeton University Press.

QUENTIN A., GUÉRIN I., 2013, « La randomisation à l'épreuve du terrain. L'exemple du projet de microassurance SKY au Cambodge », *Revue Tiers Monde*, *1*, 213, p. 179‑200.

RAVALLION M., 2009, « Should the Randomistas Rule? », *The Economists' Voice*, *6*, 2.

REDDY S.G., 2012, « Randomise This! On Poor Economics », *Review of Agrarian Studies*, *2*, 2, p. 60‑73.

RODRIK D., 2008, « The new development economics: we shall experiment, but how shall we learn? », *HKS Working Paper*, *55*.