DOCUMENT DE TRAVAIL DT/2019-04

Lies, damned lies, and RCT : une expérience de J-PAL sur le microcrédit rural au Maroc

Florent BEDECARRATS

Isabelle GUERIN

Solène MORVANT-ROUX

François ROUBAUD

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

Lies, damned lies, and RCT : une expérience de J-PAL sur le microcrédit rural au Maroc

Florent Bédécarrats Division des Evaluations de l'AFD 5, rue Roland Barthes, 75598 Paris, France bedecarratsf@afd.fr	Isabelle Guérin IRD, UMR CESSMA, 75013 Paris, France UMR CESSMA, 75013 Paris, France isabelle.guerin@ird.fr
Solène Morvant-Roux Graduate School of Social Sciences UNIGE-G3S, University of Geneva, Switzerland Solene.Morvant@unige.ch	François Roubaud IRD, UMR DIAL, 75010 Paris PSL, Université Paris-Dauphine, LEDa, UMR DIAL, 75016 Paris, France roubaud@dial.prd.fr

Document de travail UMR DIAL

Février 2019

Résumé

Comment expliquer le succès académique d'une étude randomisée dont la validité, tant interne qu'externe, est pourtant très problématique ? Prenant l'exemple d'une étude menée par le laboratoire J-PAL sur le microcrédit rural marocain, cet article mobilise les outils analytiques de la statistique, de l'économie politique et de la sociologie des sciences pour répondre à cette question. Il décrit l'ensemble de la chaîne de production de l'étude, depuis l'échantillonnage jusqu'à la publication et la dissémination des résultats, en passant par la collecte de données, la saisie et le recodage, les estimations et les interprétations. Il met en évidence une stratégie particulièrement offensive qui permet aux chercheurs de J-PAL de faire table rase du passé, y compris en s'affranchissant d'une « culture de la donnée », de refuser la critique et de contourner les règles de base de l'exercice scientifique tout au long du processus de recherche. Bien au-delà de J-PAL, nos analyses questionnent la supposée supériorité des méthodes randomisées tout en reflétant un malaise grandissant au sein du champ académique, qui parvient de moins en moins à faire respecter les règles de base de l'éthique et de la déontologie scientifique.

Mots-clefs : Randomized Control Trial (RCT), microcrédit, Réplication, Maroc, validité interne, validité externe, sociologie des sciences

Abstract

How can we explain the academic success of a randomized study whose validity, both internal and external, is very problematic? Drawing on a study conducted on Moroccan rural microcredit by J-PAL, this article uses analytical tools from statistics, political economy and sociology of science to answer this question. It describes the entire study production chain, from sampling, data collection, data entry and recoding, estimates and interpretations to publication and dissemination of results. It highlights a particularly aggressive strategy carried out throughout the study process and in the field of research. This allows J-PAL researchers to put the past behind them, including by freeing themselves from a "data culture", rejecting criticism and bypassing the basic rules of scientific exercise throughout the research process. Well beyond J-PAL, our analyses question the supposed superiority of randomized methods while reflecting a growing unease within the academic field, which is less and less successful in enforcing the basic rules of ethics and scientific deontology.

Keywords: Randomized Control Trial (RCT), microcredit, Replication, Morocco, internal validity, internal validity, sociology of sciences

JEL Codes: A11, A14, B41, C18, C93, N27, O16.

Introduction

Depuis maintenant près de deux décennies, un vent nouveau souffle dans le champ de l'économie du développement : les expérimentations randomisées, plus connues sous leur acronyme anglais de RCT : *randomized control trials*. Initiée au début des années 2000, cette approche a connu un succès fulgurant, au point de devenir une vague mondiale, non seulement en matière de recherche mais plus largement dans l'ensemble de la communauté internationale traitant des questions de développement. Sur le premier front, les RCT sont aujourd'hui largement considérées comme la référence, l'étalon-or (le *Gold standard*), dans le domaine des évaluations d'impact. Elles y exercent une double domination. Théorique d'abord, puisqu'elles prétendent être les seules à même de résoudre « rigoureusement », mais aussi de surcroit simplement, le casse-tête de l'inférence causale, devenu au fil des ans le sujet central et l'horizon indépassable des économètres (Cartwright, 2007; Jamison, 2017), soit le fameux « *experimental ideal* » (Angrist et Pischke, 2009) ; empirique ensuite, car fortes de cette qualité, les expériences randomisées sont devenues le seul instrument crédible permettant de prouver « scientifiquement » l'effet réel de politiques publiques et de programmes mis œuvre sur le terrain, et partant de sélectionner ceux qui marchent et donc susceptibles d'être répliqués partout dans le monde. Fortes de ces revendications, et surfant sur un contexte porteur (celui de l'*evidence based policy*), les RCT ont séduit un nombre croissant de bailleurs de fonds internationaux (publics et privés), en mal de prouver l'efficacité de l'aide au développement. Non seulement les RCT se sont imposées au Sud, mais par effet boomerang elles ont également donné un nouvel impetus aux évaluations d'impact dans les pays du Nord, en particulier dans des pays où la culture de l'évaluation scientifique des politiques publiques était peu présente, comme par exemple en France, où le terme « d'expérimentation sociale » est devenu en quelques années un maître-mot.

Portées par un mouvement pro-RCT particulièrement actif et puissant, des centaines de RCT ont été conduites, mobilisant des millions de dollars et conduisant à des centaines de publications dans des revues scientifiques prestigieuses (Bédécarrats *et alii*, 2018). Ce succès s'est accompagné d'un certain nombre de critiques quant à la portée universelle de la méthode et ses applications (voir par exemple pour les plus fameuses Heckman, 1991 ; Rodrik, 2008 ; Ravallion, 2009 ; Barrett and Carter, 2010 ; Deaton, 2010 ; Harrison, 2011 ; Deaton and Cartwright, 2018) y compris nos propres travaux (Bédécarrats *et alii*, 2013 et 2018). Les principaux points d'interrogation portent sur la validité interne et surtout externe de la méthodologie, avec comme conséquence sa capacité à tirer des enseignements généraux en matière de politique de développement (*what works and what doesn't work*), une revendication centrale du mouvement (Ogden, 2017) ; plus marginalement, certains travaux ont remis en cause l'aspect novateur de la méthode dans une perspective historique de la randomisation (Oakley, 2000 ; Jamison, 2017 ; Labrousse, 2017), tandis que d'autres s'attelaient à décrypter l'économie politique de ce succès (Jatteau, 2016 ; Bédécarrats *et alii*, 2018). Néanmoins, et en dépit de la renommée internationale de certaines de ces voix critiques, le mouvement pro-RCT a poursuivi sans ciller sa route victorieuse pour imposer cette nouvelle norme, au point qu'il n'est pas illégitime d'affirmer qu'il semble avoir gagné, au moins provisoirement, la bataille des idées dans ce domaine.

Si dans l'ensemble, les promoteurs des RCT n'ont pas changé leur discours quant à la supériorité intrinsèque de la méthode sur toutes les autres, et poursuivi leur politique de promotion multi-cibles et particulièrement efficace, certaines pratiques ont été amendées à la marge. De ce point de vue, le numéro spécial de la revue l'*American Economic Journal: Applied Economics*, publié en 2015 et consacré au microcrédit est emblématique (Banerjee *et alii*, 2015a). En dehors de traiter d'un des sujets majeurs de lutte contre la pauvreté, ce numéro marque une inflexion en matière de bonnes pratiques, que les détracteurs du mouvement pro-RCT leur avaient reproché de ne pas respecter. Pour la première fois, le numéro dans son ensemble, et son introduction générale en particulier, apportent des éléments de réponse à cinq types de critiques récurrentes adressées au mouvement pro-RCT (Bédécarrats *et alii*, 2018) : il développe un modèle théorique, pour répondre aux critiques

« d'empirisme agnostique » faite aux expérimentations aléatoires ; une estimation coût-bénéfice (*cost effectivness*) est proposée pour montrer la validité de l'approche au-delà du simple impact causal ; les questions de sélection au traitement (taux de *take up* en l'occurrence), de précision des estimateurs et de l'hétérogénéité du traitement sont reconnues et discutées (validité interne) ; le rôle des contextes, appréhendés à travers la diversité des terrains, des produits et des institutions des six articles permet aux éditeurs du numéro d'affirmer qu'il traite un « *échantillon raisonnablement représentatif de l'industrie/mouvement du microcrédit à l'échelle mondiale* » (Banerjee *et alii*, 2015a, p. 2 ; validité externe) ; enfin, le numéro est censé fournir les bases de données originales, pour répondre au grief d'opacité et faciliter les méta-analyses, une avancée décisive en matière de transparence et d'éthique.

Le numéro dans son ensemble a connu un succès d'estime remarquable, à la fois dans le champ de la recherche académique, puisque la prestigieuse revue *Sciences* a offert ses colonnes aux RCT (Banerjee *et alii*, 2015b)[1], mais aussi dans le champ des politiques de développement, J-PAL s'en étant inspiré pour publier une document de synthèse (*policy briefcase*) basé sur ces six articles pour tirer des conclusions générales en matière de stratégie mondiale d'accès à la finance (J-PAL et IPA, 2015).

C'est justement en s'appuyant sur le dernier point mentionné ci-dessus (la mise en ligne des bases de données) que nous avons décidé d'entreprendre le présent travail. Parmi les six articles, nous avons choisi celui de Crépon *et alii* (2015 ; dans la suite noté CDDP, initiales du nom des auteurs) avec pour objectif principal d'analyser sur un cas concret dans quelle mesure ont été prises en compte les interrogations génériques suscitées par la pratique des RCT dans le champ du développement, en appliquant comme grille d'analyse les critères énoncés ci-dessus. Plusieurs raisons déterminantes ont guidé ce choix. D'abord, ce papier qui présente les résultats d'une RCT sur l'impact de l'accès au microcrédit en milieu rural marocain auprès de l'organisation Al Amana (RCT-AAA dans la suite du texte) entre 2006 et 2010 présente un intérêt scientifique de premier plan dans le champ de la microfinance. Comme le soulignent ses auteurs, l'expérimentation est très ambitieuse dans ses objectifs, et ses propriétés exceptionnelles, aussi bien en matière de conception du protocole que de mise en œuvre opérationnelle. Sur ce premier plan, en s'intéressant aux zones rurales marocaines, elle aborde un champ relativement inexploré par les RCT, la majorité des évaluations randomisées d'impact du microcrédit étant à l'époque focalisées sur les zones urbaines ou rurales peu enclavées. De plus, grâce à une stratégie d'identification subtilement conçue, la RCT marocaine ne mesure pas seulement l'impact direct, sur les emprunteurs et les zones d'accès au traitement, mais également les externalités induites par le programme. Le choix des zones d'enquêtes semble idéal pour une RCT, puisqu'il est annoncé qu'« *elle a été réalisée dans une région où la pénétration du microcrédit en dehors de l'IMF était absolument nulle, avant ou après l'introduction du produit, et pour toute la durée de l'étude* » (CDDP, p.124 ; notre traduction ; nous reviendrons sur ce point ultérieurement). In fine, cette conjonction de facteurs favorables permet aux auteurs de conclure que « *le microcrédit apparaît comme un instrument financier puissant pour les pauvres* » (CDDP, p.148 ; notre traduction). Même s'il ne suffit pas pour sortir de la pauvreté, au moins sur le moyen terme (soit les deux années couvertes par le dispositif expérimental), les auteurs n'en excluent pas la possibilité à plus long terme, ce que devrait permettre de tester la nouvelle vague d'enquête en cours au moment de la publication de l'article[2]. L'intérêt suscité par l'article est avéré puisque, trois ans après sa publication, il fait l'objet de 248 citations (*Google Scholar ; 25/02/2019*). Il est même considéré par des figures

[1] La revue *Science* avait déjà ouvert ses colonnes aux RCT, à propos d'un projet de microfinance aux Philippines, également mise en œuvre par J-PAL (Karlan et Zinman, 2011), ainsi que l'analyse du rôle des réseaux et des mécanismes de diffusion de l'information dans l'adoption d'un programme de microcrédit en Inde (Banerjee *et alii,* 2013).

[2] Ce projet de troisième vague est d'ailleurs confirmé deux ans plus tard dans le registre de l'expérimentation, actualisé en 2017 : https://www.socialscienceregistry.org/trials/371.

majeures du mouvement pro-RCT comme une contribution décisive au débat pour comprendre l'impact du microcrédit (Ogden, 2017). En second lieu, ce papier est signé par des chercheurs du mouvement pro RCT parmi les plus prestigieux, dont Esther Duflo, sa leader médiatique, un gage attendu de qualité des analyses.

De notre côté, l'intérêt est double. D'une part, en tant que spécialistes de la microfinance, cette étude revêt une importance majeure pour tous les résultats novateurs annoncés par les auteurs. Elle l'est d'autant plus que deux d'entre nous ont mené une étude qualitative sur le même terrain (Morvant-Roux *et alii*, 2014), en complément à l'expérimentation et à la demande du commanditaire de l'expérimentation, l'Agence Française de Développement (AFD), pour laquelle travaille aujourd'hui un troisième d'entre nous. Enfin, le quatrième auteur apporte à l'équipe ses compétences en matière d'enquêtes ménages et auprès des micro-entreprises. Cet article offre donc l'occasion rêvée de poursuivre un dialogue méthodologique, entre approches qualitatives et quantitatives, malheureusement tout juste amorcé précédemment. D'autre part, l'accessibilité des micro-données ouvre la possibilité unique de se lancer dans un travail de réplication[1], une pratique courante dans le domaine des expériences randomisées conduites dans le champ médical, mais jusqu'ici impossible à mener pour les RCT réalisés par les économistes du développement, faute de mise à disposition publique des enquêtes (0,16% des RCT publiées ont fait l'objet d'un *replication test* au sens de Clemens ; Clemens, 2017 et Sukhantar, 2017).

Ce papier s'inscrit dans un projet plus large composé de deux articles complémentaires (*companion papers*). Le premier a consisté en un exercice de réplication des résultats de la RCT tel que publiés par CDDP en 2015 (Bédécarrats *et alii*, 2019). Nous y montrons que le protocole de la RCT et les données collectées sont tellement entachés de problèmes et d'erreurs de tous ordres que les conclusions qui en sont tirées questionnent aussi bien leur validité externe qu'interne. Celui-ci, le second, s'appuie sur les conclusions du premier, qui mettent en lumière les lacunes majeures de la RCT en question, pour chercher à comprendre comment de telles lacunes ont pu advenir. A travers une analyse scrupuleuse du contexte et des acteurs sur une période de près d'une décennie, nous passons au crible les dessous de la RCT en question. Du point de vue méthodologique ce papier permet d'aller beaucoup plus loin que ce qui est habituellement entrepris. En effet, outre la réplication et l'étude qualitative mentionnée ci-dessus, nous avons également mobilisé les articles académiques publiés par les chercheurs de l'AFD tirant les enseignements de leur expérience. Nous avons également eu accès à un ensemble quasi exhaustif de la littérature grise et des documents internes (notes AFD, rapports du comité de pilotage et présentations PowerPoint, rapports de suivi du projet par J-PAL, échanges de courriers électroniques). Nous avons également réalisé une série d'entretiens qualitatifs avec différents acteurs de la RCT (dirigeants et agents de crédit de l'IMF, intervenants à tous les stades du commanditaire, l'AFD, responsables et assistants de recherche de J-PAL impliqués dans le projet[2], et bien sûr enquêtés sur le terrain). Enfin, ce travail s'appuie aussi sur la méthode de l'observation participante, bien connue des anthropologues (par exemple, Olivier de Sardan, 1995). Consistant à prendre part aux activités menées par le projet, dans un rôle secondaire, elle favorise un contact rapproché et une collaboration avec les principaux acteurs. Ce poste d'observation privilégié offre une opportunité exceptionnelle pour interroger de manière pertinente

[1] Des 6 RCT présentées dans le numéro spécial, les micro-données n'étaient accessibles que pour 3 d'entre-elles.

[2] Les entretiens directs avec les responsables de la RCT et les auteurs des publications qui en ont été tirées ont été plus limités, du fait de leur réticence à engager ce type d'échanges. Néanmoins, nous pouvons saisir leur point de vue à travers leurs écrits (rapports d'étapes non publiés pour l'AFD, ainsi que l'article sous revue). La mobilisation de cette littérature grise nous a conduits à adopter un système de référencement spécifique. Les documents publiés appelés dans le corps du texte suivent la notation habituelle. Les présentations publiques sous forme de *slides* sont notées I_n°*x ;* enfin les documents interne sont notés D_n°*x*. Le tout est consigné dans la section *Références* en fin d'article.

les phénomènes que le chercheur se propose d'analyser. Cet ensemble d'informations et le suivi en temps réel et sur la durée, en général inaccessible au chercheur, est le seul moyen d'accéder au coulisses de la « science en action ». C'est d'ailleurs cet appareillage empirique qui nous a permis de mettre en pratique, dans une perspective globale de sociologie des sciences, le cadre théorique de la sociologie de la traduction (Callon, 2006), en la combinant avec des éléments (concepts de « champ » et de « capital ») empruntés à P. Bourdieu (1976). Ce travail mobilise donc les outils analytiques de la statistique, de l'économie politique, de la sociologie et de l'anthropologie du développement.

Dans cet article, nous proposons d'éclairer un triple paradoxe. Le premier est le hiatus entre une RCT entachée d'erreurs et le destin enviable du principal papier académique qui en a été tiré par les chercheurs de J-PAL. Le deuxième paradoxe provient des conclusions diamétralement opposées que tirent les deux principaux acteurs de la RCT : alors que CDDP, et plus largement J-PAL et le mouvement pro-RCT font de l'expérimentation un exemple de réussite scientifique dont on peut tirer des leçons plus générales, l'AFD, le commanditaire, en conclut qu'il faut privilégier d'autres méthodes. Enfin, le troisième paradoxe résulte du contraste entre la simplicité supposée de la méthode (quantifier l'impact causal par une simple comparaison de moyenne entre le groupe de traitement et le groupe de contrôle, grâce à l'identification d'un contrefactuel expérimental), qui constitue un des arguments fondateurs mis en avant par ses promoteurs pour affirmer la supériorité générique des RCT sur toute autre méthode, et le protocole effectivement appliqué par la RCT marocaine est d'une complexité insoupçonnée.

Le plan de cet article s'organise comme suit. Après avoir présenté les grandes lignes de l'expérimentation, la première partie expose le premier paradoxe que nous cherchons à expliquer. Il s'appuie notamment sur la synthèse des principaux résultats de notre réplication. La deuxième partie décrit finement le protocole technique et institutionnel de la RCT, les ajustements et les « bricolages » opérés en cours de route, mettant en lumière les jeux d'acteurs en présence et la distance entre les engagements théoriques annoncés *ex ante* (simplicité de la méthode) et réitérés *ex post*, et la vérité du terrain (protocole extraordinairement complexe ; notre troisième paradoxe). Dans la troisième partie nous avançons des éléments d'interprétation des principales défaillances mises en lumière dans les deux sections précédentes. Plus largement, nous offrons une relecture de l'ensemble du processus qui permet de lever les trois paradoxes énoncés ci-dessus. Enfin, nous tirons en conclusion les enseignements plus généraux de notre travail.

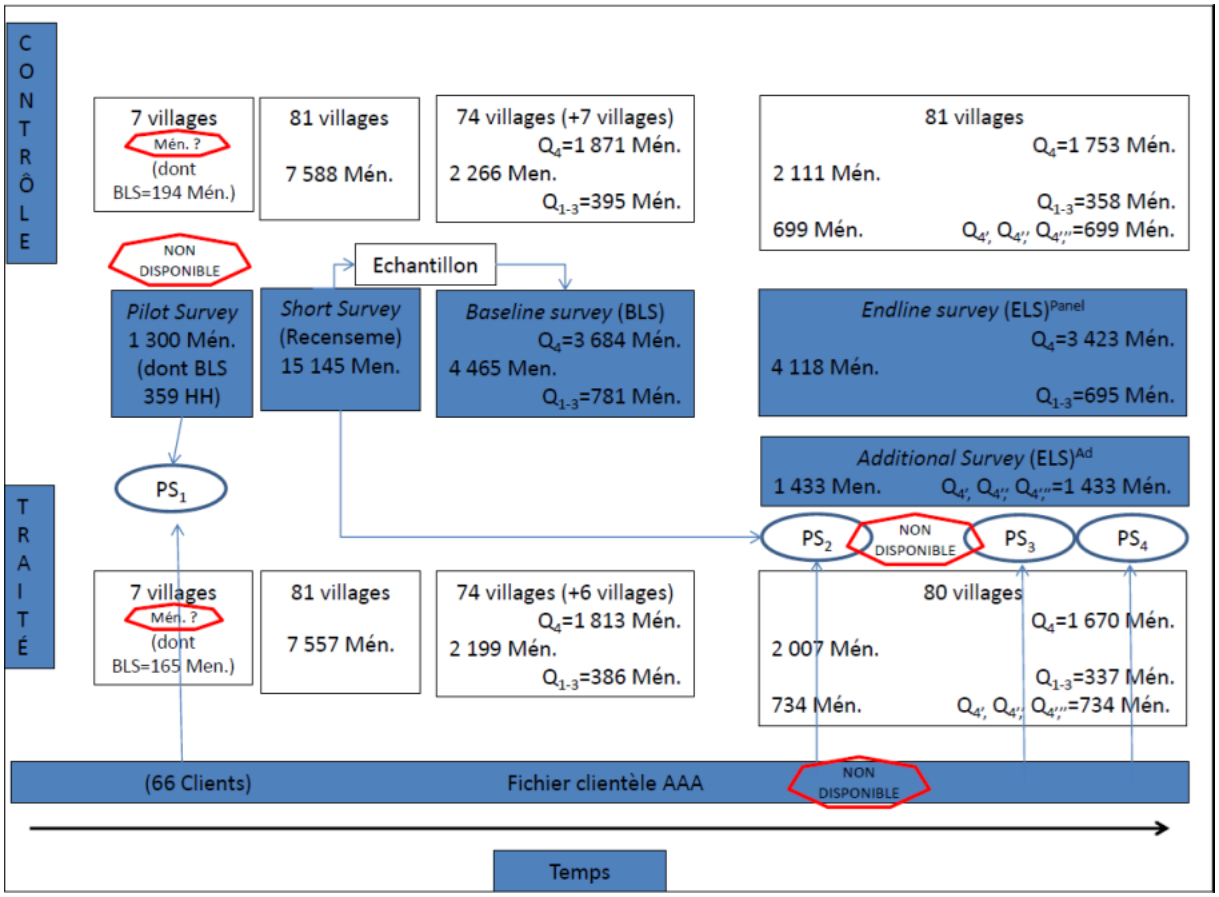
I.- Retour sur un triple paradoxe

Entre 2006 et 2010, le laboratoire J-PAL a conduit une RCT en milieu rural marocain. Il s'agissait de mesurer l'impact du microcrédit de la plus importante institution de microfinance (IMF) l'Association Al Amana (AAA dans la suite du texte), qui désirait s'implanter au-delà des zones urbaines où jusque-là elle avait concentré ses interventions, afin de mailler l'ensemble du territoire national. AAA ayant déjà amorcé un début d'expansion en milieu rural, la RCT se focalise sur les zones « enclavées ».

Le protocole proposé par J-PAL suit en apparence les standards canoniques en matière d'expérimentation aléatoire. 162 villages ont été tirés au hasard dans la zone où AAA projetait de s'implanter, au sein desquels 81 paires de villages ont été constituées, sur la base de variables observables enregistrées dans une enquête préalable à la RCT proprement dite. Pour chaque paire, un village a été assigné au groupe de traitement et l'autre au groupe de contrôle, toujours de manière aléatoire. Dans les 81 villages de traitement, AAA a démarré ses opérations, en ouvrant des antennes et en proposant des produits similaires à ceux offerts en milieu urbain (prêts solidaires puis individuels). Dans chacun des 162 villages, une enquête préparatoire et légère a été appliquée soit à l'ensemble des ménages pour les villages comportant jusqu'à 100 ménages soit à 100 ménages aléatoirement sélectionnés pour les villages plus grands (au total, 15 145 ménages et 25 variables). Cette population a ensuite été scindée en deux groupes : les ménages dits « à haute propension à emprunter », correspondant au dernier quartile d'un score de propension estimé à partir de

l'enquête préparatoire, et les autres. Tous les ménages appartenant au premier groupe ont été inclus dans l'échantillon, ainsi que 5 ménages appartenant au second groupe[1]. Deux enquêtes ont été conduites auprès de ces ménages : la *baseline*, en quatre vagues successives entre 2006 et 2007, avant l'intervention d'AAA, puis 2 ans plus tard, la *endline*. A cet échantillon de 4 465, a été ajouté un échantillon additionnel de 1 433 ménages « à très haute propension à emprunter », pour faire face au faible taux de *take up* observé en cours de route. Ces derniers ont été tirés parmi les ménages ayant fait l'objet de l'enquête préparatoire légère, à partir du calcul d'un nouveau score de propension mieux ajusté car calculé à partir de la connaissance des ménages qui avaient effectivement pris du microcrédit entre la *baseline* et la *endline*. Les nouveaux ménages ont été tirés parmi ceux qui n'avaient pas été enquêtés en *baseline*. Ils n'ont pu faire l'objet d'une enquête complète qu'en *endline*. Dans cette enquête, les 3 échantillons ont été mixés après calcul de nouveaux coefficients d'extrapolation pour assurer la représentativité au niveau de chaque village. Le score de propension a été utilisé comme instrument pour apprécier, au-delà de l'*Intention to Treat* et du *Treatment on the Treated*, le *Local Average Treatment Effect*. Cette spécificité est d'ailleurs présentée comme une contribution essentielle et inédite de cette expérimentation. La figure 1 synthétise les différentes étapes et composantes du protocole de la RCT marocaine.

Figure 1 : le dispositif de la RCT-AAA



Sources : les auteurs

Note : BLS = *BaseLine* Survey ; ELS = *EndLine* Survey ; Mén. : Ménages ; PS_n = Score de propension à emprunter, estimés à différentes étapes du protocole. Q_1-Q_4 = quartiles des propensions estimées (Q_4 le plus élevé). « Non disponible » = non fournis dans les données en ligne. La distribution des 1 300 ménages de la *Pilot Survey*, entre traités et contrôle, n'est pas connue.

[1] L'échantillon effectif est inférieur à l'échantillon théorique (4 860), du fait que certains villages comptaient moins de 100 ménages.

Les auteurs mettent en avant comme principaux résultats de la RCT que le programme n'a eu aucun impact sur la création de micro-entreprises, mais qu'il a stimulé l'activité des entreprises existantes, principalement agricoles (revenu, capital, investissement et profits). En revanche, ni le revenu des ménages, ni la consommation ne connaissent d'amélioration, du fait plus particulièrement de la réduction des revenus du travail salarié exercé en dehors du ménage. Cependant cet impact positif est très hétérogène. Le microcrédit profite avant tout aux activités les plus rentables, l'impact étant même négatif pour les autres. De plus, le programme est aussi négatif sur le taux de salariat hors du ménage et les salaires correspondants, et non significatif sur la consommation des ménages. Enfin, il ne semble pas jouer sur l'*empowerment* des femmes au sein du ménage, ni générer d'externalités importantes. Finalement, la principale conclusion de l'étude est que l'impact global du microcrédit est limité et qu'il convient de ne pas le sur-estimer[1]. De plus, sur le plan méthodologique, les auteurs mettent en avant la stratégie d'échantillonnage qu'ils ont élaborée pour palier le faible taux de *take up* observé, et qui pourrait servir de modèle pour d'autres expérimentations tant ce problème est récurrent.

IA- La RCT *Al Amana* au crible de sa réplication

La mise à disposition des micro-données combinée aux limites identifiées dans l'enquête qualitative associée (Morvant-Roux *et alii*, 2014), nous ont incités à entreprendre un exercice de réplication, dans le but de reproduire et d'analyser les résultats présentés dans leur article par CDDP ; un exercice appelé de ses vœux par un certain nombre de chercheurs (Duvendack *et alii*, 2017), mais à ce jour encore marginal dans le champ du développement. L'objectif de ce travail, dont la communauté scientifique recommande l'usage plus systématique, est de déterminer la validité des résultats empiriques de l'étude. Il est particulièrement ambitieux, dans la mesure où nous nous livrons à trois des quatre types de réplications de la typologie proposée par Clémens (2017) : la catégorie dites des *replication-test* (qui consiste à utiliser les mêmes spécifications que les auteurs sur le même échantillon pour vérifier qu'on trouve les mêmes résultats), elle-même subdivisée entre les *replication-verification* (visant à estimer les erreurs de mesure : données de base, recodage, programmation) et les *replication-reproduction* (visant à identifier les erreurs de sondage), et celle des *robustness test*, dans la sous-catégorie des *robustness reanalysis* (en modifiant les analyses à partir de variables recodifiées). Le seul type de réplication non mis en œuvre est celle des *robustness extension* qui applique les mêmes procédures à d'autres populations. Le recodage intégral des programmes écrits sous *Stata* dans le logiciel statistique *R* et l'obtention de résultats parfaitement identiques est une garantie supplémentaire de la parfaite fiabilité de notre réplication. Nous nous bornerons ici à synthétiser les résultats obtenus par Bédécarrats *et alii* (2019).

Dans l'ensemble, le diagnostic de la réplication est alarmant, tout particulièrement du fait de données de base d'une qualité exceptionnellement mauvaise, qui se combine avec des problèmes d'échantillonnage et de traitement des données, qui nous conduisent à remettre en doute l'ensemble des résultats de l'étude.

En premier lieu, très loin de la simplicité affichée comme un des atouts majeurs de la méthodologie des RCT (comparaisons de moyenne entre 2 groupes sélectionnés de manière aléatoire pour identifier un groupe de traitement et un groupe de contrôle), et dont elle tire son pouvoir de séduction, le protocole employé s'éloigne très sensiblement de ce cadre canonique. L'expérimentation proposée est d'une surprenante complexité méthodologique (voir figure 1 et infra). La conséquence en est qu'il est difficile de se faire une idée claire de ce qui est finalement démontré. En second lieu, l'ensemble du protocole empirique mis en place sur le terrain cumule un

[1] Notons néanmoins que, comme mentionné en introduction, la conclusion de l'article CDDP souligne que le "*microcredit appears to be a powerful financial instrument for the poor*" (p.148) ; une assertion pour le moins contradictoire sur laquelle nous reviendrons dans la partie 3.

nombre important de défaillances majeures, et ce à toutes les étapes du processus, depuis la conception de l'expérience (en amont), jusqu'à la présentation des résultats (en aval), en passant par la conduite des enquêtes sur le terrain et les traitements statistiques et économétriques. Dans ce contexte, on s'interrogera dans la troisième partie pourquoi des défaillances aussi rédhibitoires n'ont pas conduit à un rejet de l'article par la revue, où pour le moins à des mises en garde dans le texte pour relativiser la portée des résultats.

Développons brièvement chacun des points, qui portent aussi bien sur les erreurs de mesure que les erreurs de sondage (pour une présentation détaillée, voir Bédécarrats *et alii*, 2019) :

- la qualité des données collectées est tellement douteuse qu'on peut s'interroger de savoir si elles sont tout simplement exploitables : un nombre substantiel de ménages enquêtés dans l'enquête de base (*baseline survey*) et considérés comme ré-enquêtés dans l'enquête finale (*endline survey*), donc traités comme tels (données de panel) ne sont à l'évidence pas les mêmes. Par exemple, au moins 20% des ménages ont une composition tellement différente par sexe et âge entre les deux enquêtes qu'il n'est pas plausible qu'il s'agisse des mêmes ménages. En conséquence, les groupes de traitement et de contrôle en *endline* ne sont plus les mêmes qu'en *baseline* ;
- l'identification du groupe de traitement (appartient à un village avec une antenne de l'IMF et bénéficie d'un microcrédit de cette IMF) dans l'enquête de base est sujette à caution. Les auteurs ont préféré retenir l'information issue du fichier clientèle de l'IMF pour identifier le groupe traité, plutôt que de se baser sur les enquêtes (*baseline survey* et *endline survey*), tout en conservant toutes les autres informations issues des enquêtes telles quelles pour l'analyse, alors qu'aucune raison ne permet de penser qu'elles sont plus fiables. Les données du fichier clientèle sont elles-mêmes douteuses. Ainsi, 152 des 346 ménages (44%) qui déclarent avoir obtenu du microcrédit de l'IMF en *endline* n'apparaissent pas dans le fichier clients. A contrario, 241 des 435 ménages (55%) qui apparaissent dans le fichier clients déclarent ne pas avoir souscrit de crédit auprès de l'IMF ;
- contrairement à ce qui est annoncé dans l'article, la zone d'enquête n'est pas du tout vierge d'accès au crédit en *baseline*[1]. 11% des 2 007 ménages du groupe de traitement avaient accès à une forme de crédit formel en *baseline.* Ce taux est resté constant en *endline*. On observe un simple effet de substitution entre le microcrédit proposé par l'IMF et les autres sources de crédit, ce qui conduit à s'interroger sur le type d'impact qu'on estime in fine ;
- les recodages des variables de base des enquêtes, presque toujours menés sans discussion critique, sont dans de nombreux cas erronés. C'est notamment le cas pour les différentes sources de crédit, pourtant tout à fait centrales puisqu'il s'agit de la variable de traitement. On observe également des erreurs de mesure massives sur les variables d'*outcomes* (équipement des ménages - *assets* -, revenus, emplois) et les variables de contrôle, ainsi que des variations erratiques entre les deux enquêtes (par exemple, tracteurs et moissonneuses – les actifs de plus forte valeurs observés dans l'étude – ont été comptabilisés en *baseline* et pas en *endline* ; les prêts émis par d'autres IMF ont été pris en compte dans l'accès au crédit en *baseline* et pas en *endline*). En fait, c'est la plupart des variables d'intérêt des deux enquêtes qui sont sujettes à caution ;
- du point de vue des échantillons enquêtés, le protocole est entaché d'erreurs de sondage : d'une part, il n'est pas possible de savoir de quelle population ils sont censés être représentatifs : les ménages marocains (qui ont pris du microcrédit), comme annoncé dans le

[1] "*The study has three features that make it a good complement to existing papers. First, it takes place in an area where there is absolutely no other microcredit penetration, before or after the introduction of the product, and for the duration of the study.*" (CDDP, p.124)

titre de l'article ? L'ensemble des ménages ruraux marocains (voir CDDP, abstract)[1] ? Les ménages ruraux marocains des zones déshérités (voir CDDP, conclusion) ? Les ménages ruraux marocains des zones déshéritées où l'IMF a commencé à intervenir grâce à la RCT ? Les ménages ruraux marocains des zones déshéritées où l'IMF a commencé à intervenir grâce à la RCT et susceptibles de prendre du microcrédit ? Les ménages ruraux marocains des zones déshéritées où l'IMF a commencé à intervenir grâce à la RCT et susceptibles de prendre du microcrédit, une condition identifiée par l'estimation d'un score de propension *ex ante* (dont le pouvoir explicatif est très faible) ? Les ménages ruraux marocains des zones déshéritées où l'IMF a commencé à intervenir grâce à la RCT et susceptibles de prendre du microcrédit, condition identifiée par l'estimation d'un score de propension amélioré *ex post* (dont le pouvoir explicatif est également très faible et inconsistant avec le score *ex ante*) ? Au mieux, du point de vue du sondage, l'échantillon ne vaut que pour lui-même et les résultats de la RCT n'autorisent aucune extrapolation à quelque sur-population que ce soit ;

- les scores de propension successifs estimés pour identifier la population centrale d'intérêt du papier (les ménages à « haute propension à emprunter ») ne permettent pas de prédire la probabilité d'emprunt des ménages. Ils sont par ailleurs incohérents entre eux. Enfin l'ajout d'un échantillon de ménages enquêtés en *endline* mais pas en *baseline*, pour compenser le faible taux de *take up*, observé, remet en question la représentativité de l'échantillon, même pour couvrir les 162 villages enquêtés (probabilité d'inclusion variable en cours d'expérimentation) ;
- les échantillons de traitement et de contrôle ne sont pas équilibrés en *baseline*. Pour ne prendre que quelques exemples (non mentionnés dans l'article de CDDP) en ne retenant que les différences significatives, l'écart entre groupe de contrôle et groupe de traitement est de 21% pour la proportion de ménages ayant emprunté (et même de 50% pour les crédits formels hors AAA), de 9% pour le montant des biens durables (*assets*), et de -11% pour la production et les dépenses, au détriment cette fois du groupe traité. Ils sont donc incomparables, contrairement à ce qu'affirment les auteurs, qui ne présentent des tests d'équilibre en *baseline* que sur des variables d'une importance secondaire. Ces dernières ont été sélectionnées de manière *ad hoc*, en ne retenant que les variables observables les plus proches en moyenne entre les deux groupes, alors que nombre d'autres présentent des différences significatives ;
- la procédure de *trimming* des observations aberrantes (*outliers*) est inconsistante, *ad hoc* et différente entre la *baseline* et la *endline* alors que les estimations sont fortement dépendantes de l'exclusion des *outliers* ;

Au total, les résultats de la RCT doivent être remis en question aussi bien du point de vue de sa validité interne de la RCT, que de sa validité externe. Leur pouvoir d'extrapolation, leur champ d'inférence et de généralisation ne garantit même pas leur représentativité au niveau des villages enquêtés.

IB- Le fabuleux destin d'un article contestable

En dépit de ses lacunes majeures, la RCT-AAA a connu un grand succès aussi bien dans le champ académique qu'au sein de la communauté du développement : c'est là le premier paradoxe. Du côté de la recherche, après trois ans, le papier a été cité 248 fois (Google Scholar, 25/02/2019)[2]. Ce chiffre, très conséquent en soi, occupe aussi une position enviable en termes relatifs. Il est l'un des

[1] Et comme le laissait penser le titre d'un premier document de travail du MIT publié en 2011 (Crépon *et alii*, 2011).
[2] Le document de travail du MIT de 2011 mentionné plus haut en compte 246 (février 2019).

plus élevé des six articles sur le microcrédit dans le numéro spécial consacré à ce sujet par la revue AEJ: AE, certes loin derrière celui de Banerjee *et alii* (2015c) et ses 51 versions sur un programme de microfinance en Inde (1 652 occurrences) mais à parité avec celui de Angelucci *et alii* (2015) sur l'IMF *Compartamos Banco* au Mexique (261), et devant les trois autres : Attanasio *et alii* (2015) sur la Mongolie en compte 190 ; Augsburg *et alii* (2015) sur la Bosnie 179 et enfin Tarozzi *et alii* (2015) en Ethiopie 170. Si le nombre de citations est censé être guidé par la portée scientifique des résultats, de cette hiérarchie, on peut légitimement induire que le prestige des auteurs compte également. A titre de comparaison, le décompte de l'article de Pitt et Khandker (1998), identifié par Roodman et Morduch (2014) comme le plus cité de tous les articles empiriques portant sur un projet individuel de microfinance, atteignait 890 occurrences sur *Google Scholar*. Mais cette performance n'était atteinte que 15 ans après sa parution[1]. Pour se faire une idée en dehors du champ de la microfinance, CDDP enregistre presqu'autant de citations que l'article vocal de Duflo *et alii* (2015) sur la gouvernance scolaire et les primes aux enseignants au Kenya, publié dans la prestigieuse revue *Journal of Public Economics* la même année (254 citations en février 2019).

Ce point est particulièrement important, car l'impact académique de l'article de CDDP dépasse largement le papier lui-même et ses lecteurs. Dans le cas d'espèce, cet effet en cascade diffuse à trois niveaux, en cercles concentriques, partiellement emboités. En premier lieu, ses résultats se répandent à travers les citations qu'il suscite. L'analyse des 248 citations montre que ce sont souvent des citations croisées des auteurs eux-mêmes, à l'origine de près de 10% d'entre elles (20 cas) ; et bien plus si on élargit le champ à l'ensemble des chercheurs affiliés à J-PAL.

En deuxième lieu, en tant que composante du numéro spécial d'AEJ : AE, dont il constitue un des piliers, il est utilisé par les auteurs eux-mêmes et leurs collègues proches pour tirer des enseignements plus généraux sur la microfinance en général, notamment dans l'introduction au numéro (Barnerjee *et alii*, 2015a). Cette dernière fait l'objet de 439 citations (février 2019). Comme mentionné dans le texte, les articles ont été spécialement calibrés, à la demande de l'éditrice en chef de la revue, Esther Duflo, pour suivre un plan commun, afin de faciliter la lecture et les enseignements comparatifs. L'article de CDDP alimente et étaye les thèses développées dans cette tentative de montée en généralité. Son rôle est crucial pour deux raisons principales : d'une part, il couvre une aire géographique différente des autres papiers, ce qui permet aux éditeurs scientifiques d'affirmer que leurs résultats combinés sont « *fairly representative of the microcredit industry/movement worldwide* » (assez représentatif de l'industrie/mouvement du microcrédit dans le monde ; notre traduction) ; d'autre part, il est l'un des seuls à porter sur le milieu rural « enclavé », une aire géographique concentrant les poches de pauvreté les plus aigües mais très peu explorée dans la littérature couverte par les RCT, et par là susceptible d'étendre encore le champ de validité supposée des conclusions tirées. La RCT marocaine est mobilisée de deux façons dans l'introduction. D'une part, elle sert à confirmer certains enseignements généraux, pour soutenir différentes thèses : le faible taux de *take-up* et les difficultés à le prédire, le faible impact « transformateur » (*transformative*) du microcrédit sur les ménages emprunteurs, mais dans un contexte de forte hétérogénéité inobservable, qui pourrait appuyer l'idée d'un impact très significatif sur certaines populations, et nul voire négatif sur d'autres (un résultat propre à la RCT marocaine). En revanche, l'impact sur les entreprises est plutôt positif (niveau d'activité, profits). D'autre part, le cas du Maroc est cité en exemple pour un certain nombre de qualités propres qui la distingue des autres expérimentations. Plusieurs de ces bonnes propriétés sont mises en avant. D'abord, la faiblesse systématique du taux de *take up* pose un problème de pouvoir statistique des RCT dans ce domaine. Les auteurs de l'introduction proposent des pistes pour lever cette contrainte. En particulier, elle pourrait être résolue par des tailles d'échantillon plus grandes, et des plans de sondage innovants / *design ad hoc*. Le cas du Maroc est cité en exemple, à côté de la RCT indienne. Un des atouts mis en exergue est également que l'expérimentation soit considérée comme ayant été menée sur un terrain

[1] Au 24 février 2019, le compteur est passé à 1 859 citations.

absolument vierge d'alternative de crédit, ce qui permet d'exclure les potentiels effets de substitution entre le traitement et des sources alternatives de financement. Ensuite, Banerjee *et alii* (2015a) loue la RCT AAA pour sa méthode originale de prédiction des taux de *take up*, ainsi que de l'estimation des effets de *spillovers*. Enfin, en termes d'indicateurs, elle fait partie des quatre qui ont estimé les dépenses de consommation des ménages et la seule à avoir mesuré l'ensemble du revenu, et pas seulement certaines de ses composantes.

Finalement, la RCT-AAA vient confirmer et donc alimenter toutes les conclusions générales tirées ci-dessus, et qui sont censées arrêter l'état de l'art dans le monde en développement sur les effets causaux du microcrédit. Ainsi, sur certains points, la RCT-AAA est louée pour avoir repoussé les frontières de la connaissance scientifique sur le microcrédit.

Enfin et en troisième lieu, le dernier cercle de cette montée en généralité est l'article publié dans *Science* la même année (Banerjee *et alii*, 2015b)[1]. Ce dernier fait référence au papier de CDDP, ainsi qu'à l'introduction du numéro spécial d'AEJ : AE, mentionné ci-dessus et qui lui-même fait appel à CDDP. Bien que cité « seulement » 382 fois (février 2019), cet article est particulièrement ambitieux. Il se propose, à partir de six RCT similaires conduites par J-PAL, de montrer qu'un *big push*, sous la forme d'un programme intégré, combinant plusieurs dimensions (subvention aux équipements productifs et à la consommation, formation, accès à des instruments d'épargne, information, etc.) et susceptible d'avoir un impact significatif sur les plus pauvres. L'article de CDDP, au même titre que les cinq autres du numéro d'AEJ: AE, sert aux auteurs à justifier l'idée que le microcrédit, s'il permet parfois d'accroître le niveau d'activité des entreprises, n'a pas d'impact sur la consommation et donc n'est pas un instrument décisif pour sortir de la pauvreté.

En fait, l'impact de l'étude de CDDP va bien au-delà des trois cercles mentionnés ci-dessus. D'abord, dans le champ académique nous n'avons relevé que les références directes à l'article. Cet ensemble ne tient pas compte des impacts indirects de second tour qui élargissent sensiblement son périmètre d'influence. En effet, tous ces articles, sont eux-mêmes cités par d'autres, et ainsi de suite (les citations des citations). Ensuite, l'article a largement inspiré la littérature non académique que nous explorerons dans la troisième partie. Bien sûr, confiants dans le processus de revue par les pairs, tous ces auteurs ne sont en rien responsables d'avoir contribué à diffuser des résultats qui ne le méritaient pas, et que notre réplication a permis d'identifier. La question à ce stade est de comprendre comment un tel grand écart a-t-il pu advenir, en commençant en amont, par l'analyse fine du processus de construction et de réalisation de la RCT-AAA sur le terrain.

II. Les coulisses de la collecte de données : une étude réellement réalisée dans les règles de l'art ?

La sociologie des sciences montre que la production scientifique, y compris dans les sciences réputées les plus « dures » (les sciences de la vie et de la matière), est indissociable de dynamiques sociopolitiques multiples qui interviennent à toutes les étapes de la chaîne de production du savoir et qui mettent en jeu une grande diversité d'acteurs (dont certains ne sont souvent pas mentionnés alors que leur rôle est central), et avec eux des logiques, des intérêts et des contraintes également multiples et parfois divergentes. Explorer les coulisses de la collecte de données est donc crucial pour comprendre le décalage entre protocoles officiels, tels que présentés dans les publications des chercheurs, et ceux effectivement mis en œuvre sur le terrain.

Ici en l'occurrence, le contexte tant international qu'institutionnel de cette étude était sur le papier idéal et la rendait particulièrement prometteuse (partie IIA). L'étude se déploie dans un environnement international a priori très favorable aux RCT et au microcrédit. D'une part, le début

[1] Ce n'est pas la première fois que la revue *Science* ouvrait ses colonnes aux RCT, sur le microcrédit (Karlan et Zinman, 2011)

des années 2000 marque une période de grâce pour le microcrédit qui bénéficie d'une aura particulière parmi les acteurs influents du champ du développement. D'autre part, l'accent mis sur les décisions politiques « informées » ou *evidence based policy* renforce à cette époque l'intérêt pour les études d'impact expérimentales, les RCT prenant leur envol et jouissant d'un succès d'estime considérable (Ravallion, 2009 ; Bédécarrats *et alii*, 2018 ; Donovan, 2018).

Plus spécifiquement dans notre cas, l'étude est portée par des partenaires considérés comme des leaders de premier plan dans leur domaine (laboratoire de recherche, bailleur, institution de microfinance) et motivés à s'investir dans cette évaluation pour différentes raisons institutionnelles. Les conditions sont *a priori* réunies et conformes aux exigences du protocole théorique des RCT. Or l'analyse des coulisses de la mise en œuvre (partie IIB) met en scène un plus grand nombre d'acteurs, à commencer par les enquêtés et leurs familles mais aussi des organisations de microcrédit concurrentes, des agents de crédit, un bureau d'étude, des enquêteurs, du personnel en charge de la saisie. Chacun à leur manière, ces différents acteurs vont contribuer à distordre le protocole. Or certaines de ces distorsions, qui in fine remettent en question la validité des résultats obtenus, auraient largement pu être anticipées et contrôlées, tandis que d'autres, notamment le manque de participation, auraient mérité un minimum de réflexivité. Cette exploration permet de lever le troisième des paradoxes mentionnés en introduction : du fait de cet enchevêtrement de dynamiques, comment une méthode prétendument simple se traduit finalement par un protocole à la fois d'une extrême complexité et de surcroît opaque.

IIA. Un contexte international et un terrain idéal pour une étude qui doit faire date

Un environnement propice à la condition d'une RCT sur le microcrédit

En premier lieu, le début des années 2000 marque une période de grâce pour le microcrédit qui bénéficie d'une aura particulière dans le champ des politiques de développement articulant objectifs de lutte contre la pauvreté (OMD) et approches pro-marché. Cette période d'euphorie culmine en 2005 et 2006 avec l'année internationale du microcrédit à l'ONU et l'obtention du prix Nobel de la Paix décerné par M. Yunus et la Grameen Bank. Alors que le microcrédit est en pleine expansion à l'échelle mondiale et attire attentions et financements, pour son rôle majeur dans la lutte contre la pauvreté, sur le plan méthodologique les différentes approches peinent néanmoins à démontrer son impact. Cette question devient une préoccupation centrale pour les acteurs du secteur (Cheston et Reed, 1999). Si certains se satisfont de la preuve de l'impact positif du fait de l'existence d'une demande (de microcrédit), le mouvement des *evidence based public policies* émerge en lien avec la déclaration de Paris sur l'efficacité de l'aide et dans son sillage les acteurs du développement voudront prendre le tournant de la « rigueur » dans les évaluations d'impact. Les dissensions émergent autour d'une étude d'impact de la Grameen Bank conduite par la Banque mondiale dont la publication donne lieu à des débats animés quant aux limites des protocoles de type quasi-expérimental pour établir l'impact réel du microcrédit (Pitt et Khandker 1998 ; Morduch 1998 ; Chemin 2008 ; Duvendack *et* Palmer-Jones, 2012). Ce débat s'est prolongé durant près de deux décennies (Roodman et Morduch, 2014 ; Pitt 2014).

En second lieu, sur le papier, le contexte sur le terrain semble idéal eu égard aux multiples exigences de protocole des RCT. D'abord, la contrainte de pouvoir comparer des populations identiques dont la « seule » différence tiendrait à l'accès ou pas au microcrédit semble en théorie pouvoir être respectée. En effet, à cette époque, le microcrédit marocain est majoritairement implanté dans les zones urbaines (Reille, 2009), ce qui laisse présager que les zones rurales les plus enclavées sont vierges de toute intervention. Ce souci de l'équipe de chercheurs est efficacement relayé par le

directeur d'AAA, qui, en tant que Président de la Fédération Nationale des Acteurs du Microcrédit (FNAM), parvient à convaincre, au démarrage et tout au long de l'étude, les principaux concurrents du secteur à coopérer pour préserver la virginité des sites de contrôle. En 2007, Bruno Crépon affirmait ainsi que « *le programme est la seule différence entre les deux populations* » ce qui doit permettre la « *mise en place d'un dispositif qui produira des résultats qui feront date* » (Crépon, 2007 : 22, I_7). Ensuite, au vu de la faible présence (supposée) du microcrédit dans les zones rurales enclavées, l'équipe de recherche dispose d'un large échantillon d'unités d'observation ce qui ouvre des possibilités importantes pour réaliser au mieux le choix aléatoire et l'appariement des sites inclus dans les échantillons de traitement et de contrôle. Pourtant, en dépit d'un contexte largement propice et alors que le succès de l'étude repose "*sur le respect scrupuleux d'un protocole transparent*" (Crépon, 2007 : I_7), le protocole dérape rapidement en raison de multiples problèmes à différentes étapes de l'expérimentation.

Des institutions de premier plan pleinement engagées pour la réussite de la RCT

La RCT-AAA est proposée et conduite par le laboratoire J-PAL, qui ambitionne de développer sa méthode tous azimuts dans le champ du développement. Il se tourne donc vers différents acteurs (ONG, gouvernements et bailleurs de fonds) pour obtenir des financements. Cette expérimentation au Maroc est portée par des institutions académiques et des chercheur(e)s prestigieux(ses), de renommée internationale et principaux membres actifs de J-PAL, qui, s'il n'est pas encore la firme globalisée qu'il est devenu, est déjà le leader mondial en matière de RCT : Esther Duflo, co-fondatrice de J-PAL et économiste de renommée internationale en poste au MIT, assure la direction scientifique et organisationnelle du projet ; Bruno Crépon, interlocuteur principal et chef de projet, est quant à lui en poste au CREST/ENSAE à Paris 1, tandis que Rachel Glennerster occupe alors la fonction de directrice exécutive de J-PAL. A ces trois ténors, deux autres chercheurs complètent l'équipe de J-PAL : William Parienté, qui remplacera Rachel Glennerster et assurera la supervision du travail de terrain avec Florencia Devoto, chercheuse junior en charge (à cette époque) de la partie opérationnelle des évaluations portées par le bureau de J-PAL à Rabat.

De son côté, l'AFD est le bras armé de la France, l'un des principaux donateurs de l'APD bilatérale en matière de financement du développement. Dans le cadre de sa nouvelle politique de recherche, l'AFD cherche à développer une expertise de niveau international dans le champ de l'évaluation d'impact. Elle réalise rapidement son intérêt pour ce projet, à double titre : opérationnel et méthodologique. Sur le plan opérationnel, engagée aux côtés d'AAA depuis plusieurs années, l'AFD souhaite profiter de cette étude pour évaluer le déploiement de ce leader du microcrédit marocain vers des zones rurales reculées et contribuer à de meilleures connaissances sur le microcrédit rural, peu étudié à cette époque. Pour le département Evaluation de la Direction de la Recherche de l'AFD, cette proposition arrive à point nommé car ses responsables souhaitent alors participer à la mise en œuvre de méthodes randomisées afin de tester leur faisabilité et nourrir une réflexion méthodologique plus large visant à évaluer au mieux les effets des programmes financés par l'AFD en vue d'améliorer la redevabilité (*accountability*) de ses actions et de prendre des décisions avisées quant aux possibilités d'extension de l'intervention (*scaling up* ; Naudet *et alii*, 2012 ; I_13). Plutôt que de contribuer financièrement à de grands fonds mondiaux destinés au financement des RCT, l'AFD souhaite se positionner comme un bailleur ayant un rôle actif en matière de production de connaissances. Cette volonté de se forger une doctrine sur les apports et les limites des RCT s'inscrit donc dans une réflexion sur les différentes techniques et stratégies permettant de contribuer au

débat public, structuré autour de rencontres de bailleurs autour de ces questions, et à l'élaboration de politiques publiques. Enfin, pour l'AFD, soucieuse d'associer des partenaires opérationnels en vue de produire des résultats utiles pour l'action, le contexte marocain est idéal et les attentes de cette collaboration sur le plan opérationnel sont fortes, notamment concernant les produits de crédit mieux adaptés aux besoins des populations rurales (Delarue, 2008 ; I_8). Soulignons que ce type d'alliance, qui permet une étroite relation entre *project design* et *evaluation design,* singularise les RCT d'autres méthodologies d'évaluation d'impact notamment de type ex-post (Pamiès-Sumner, 2015) et en conditionne le succès. Le montage partenarial permet à J-PAL de faire coïncider son protocole avec le déploiement d'AAA en zones rurales, ce qui est conforme avec les exigences du protocole des RCT.

Le troisième partenaire institutionnel de la RCT est l'IMF marocaine Al Amana, classée parmi les trente « meilleures » IMF au niveau mondial (Mix-Market, 2005). La construction du secteur du microcrédit au Maroc au tournant des années 2000 implique la mobilisation de nombreux acteurs. Emanation des milieux activistes, le microcrédit acquière à cette époque sa légitimité politique et institutionnelle, bénéficiant de soutiens internationaux mais aussi d'un engagement marqué du gouvernement marocain. Les associations de microcrédit AAA et Zakoura, elle aussi dans le palmarès des trente « meilleures » IMF, s'imposent rapidement comme les leaders d'un secteur très standardisé intervenant essentiellement en milieu urbain et déployant une méthodologie identique de prêt. Dans ce contexte, l'intérêt pour cette étude est double pour AAA : alors que ses taux de croissance sont vertigineux (D'Espallier *et alii,* 2015), les 500 000 clients de cette IMF sont à cette époque essentiellement localisés dans les zones urbaines et péri-urbaines mais pas dans le milieu rural plus isolé. L'appui de l'AFD à son déploiement auquel est adossée une étude d'impact constitue une aubaine pour tester le potentiel de son expansion et réfléchir à l'adéquation des produits de microcrédit. Cela permettrait à AAA d'asseoir sa suprématie comme le souligne la responsable du développement chez AAA à cette époque : « *mailler l'ensemble du territoire était un moyen de s'imposer* » (anciennement responsable développement AAA, entretien téléphonique, 16 mars 2018). Les attentes opérationnelles concernant l'offre de services financiers à ce segment de la population constituent également, d'après notre interlocuteur[1], un moteur pour apporter un soutien étroit à cette évaluation et en garantir le succès.

IIB- La question centrale de la participation et le dérapage de la RCT

Nous détaillons dans ce qui suit les différents éléments de l'enquête qui entachent rapidement le protocole et la qualité des données collectées. Ce dernier point éclaire les multiples failles qui compromettent l'étude, soulevées par la réplication réalisée par Bédécarrats *et alii* (2019) et résumées ci-dessus dans la partie 1.

Le choix des zones enquêtées et l'épineuse question de la participation

Concernant les zones à enquêter, la méthode d'évaluation expérimentale de l'équipe de J-PAL prévoyait le choix aléatoire de 81 paires de villages situées dans des zones dans lesquelles AAA planifiait de déployer son offre de microcrédit. Deux contraintes ont orienté le choix des villages. La première est d'ordre opérationnel. Il s'agissait de ne pas perturber l'expansion d'AAA et donc de

[1] Nous employons le genre masculin pour tous nos interlocuteurs dans un souci de préserver leur anonymat.

concentrer l'étude sur les zones éloignées des nouvelles branches et les contraintes d'appariement des villages traitement et contrôle, c'est-à-dire la construction de paires de villages supposés équivalents. L'étude porte donc sur des zones peu densément peuplées et enclavées. Pourtant, si cela a pu être vrai pour les paires de villages de la première vague (c'est-à-dire la première tranche de la *baseline*), Morvant-Roux *et alii* (2014) montre que la situation a changé dès la deuxième vague d'enquêtes puisque l'échantillon des zones de J-PAL revisité par l'étude qualitative indique une diversité de contextes, allant de zones péri-urbaines à des zones rurales extrêmement enclavées (ibid : p. 307). On peut dès lors interroger la comparabilité de villages enclavés avec des villages beaucoup plus proches de centres urbains secondaires.

Une autre question centrale se pose : les zones enquêtées étaient-elles aussi « vierges » que ne l'affirme l'équipe de recherche (Crépon, 2007 ; I_7). L'AFD évoque ce point critique dans son rapport suite à la mission de suivi effectuée au mois avril 2008 en insistant sur la contamination des zones non-traitées par les IMF Ardi et Zakoura (AFD, 2008 ; DI_3). Or, si ce point fait beaucoup débat au moment de l'étude, Bédécarrats *et alii* (2019) montre que la contamination est finalement un aspect secondaire car demeurée réduite tout au long de l'étude même après avoir relaxé certaines règles d'octroi à la demande d'AAA. En effet, pour des raisons éthiques et pour ne pas pénaliser les ménages des villages de contrôle sollicitant spontanément un crédit, les équipes d'AAA ont obtenu ce changement de politique. Le problème se situait ailleurs : les villages dits traités n'ont pas subi de changement significatif par rapport à l'accès au crédit entre la *baseline* et la *endline*. Il s'est davantage agit d'un remplacement de crédits formels des autres organismes de crédit au bénéfice des microcrédits AAA. Ceci s'explique sans doute par les contraintes imposées par le directeur de l'époque aux autres acteurs du secteur pour contenir tout risque de contamination. Par ailleurs, en 2008, au beau milieu de l'enquête, le secteur du microcrédit est confronté à une grave crise d'impayés. Or celle-ci se traduit par un retrait marqué des zones rurales pour l'ensemble des organisations de microcrédit, avec un portefeuille rural qui passe de 65% à 47% du portefeuille total (Rozas, 2014). Interrogé en mars 2018, l'ancien responsable développement AAA nous confirme qu'après la crise, le risque de « contamination » s'est considérablement atténué. En définitive, la « contamination » a surtout eu lieu *avant* le démarrage de l'enquête : les zones prétendument vierges de microcrédit ne l'étaient en fait pas du fait de présence de concurrents d'AAA, qui se sont ensuite retirés.

Une autre difficulté, fort problématique, est la question du taux de *take up*. Dans les villages traités, l'équipe est confrontée à un problème qu'elle n'avait pas anticipé : celui de la participation, qui s'avère beaucoup plus faible qu'initialement envisagé. Face à ce problème, l'équipe procède à des ajustements majeurs à deux niveaux : le protocole de la RCT et les produits proposés par AAA.

Les documents de travail produits par l'équipe de J-PAL indiquent que l'équipe anticipait au départ un taux de participation très élevé dans les zones de traitement. En 2006, J-PAL mentionne en effet que « *suivant les indications de Al Amana on peut s'attendre à un très bon taux de pénétration au niveau des ménages (son expérience en milieu urbain est un taux de pénétration proche de 100%) »* (J-PAL, 2006 ; DI_1, p. 13*)*. Si J-PAL propose déjà de se concentrer sur les ménages ayant de grandes chances de prendre un microcrédit, la question de la participation réelle n'est pas soulevée. C'est dans un rapport préparé pour le comité de pilotage du mois d'août 2007 que les difficultés liées à la participation sont évoquées pour la première fois, notamment pour ce qui concerne les sites à partir de la deuxième vague d'ouverture d'antennes AAA (AFD, 2007 ; DI_2). Le rapport fait état d'une

participation plus faible que prévue, et très hétérogène : elle varie entre 0% et 55% selon les villages (*douars*). Ce dernier taux s'observe sur un seul site. Sur les 39 *douars* de traitement de la deuxième vague, 11 *douars* « *n'ont déboursé aucun crédit ou presque, 5 mois après le démarrage des activités dans la zone de traitement* » (Ibid, p. 7).

Un protocole qui déraille

Pour pallier cette faiblesse de la participation, l'équipe de J-PAL décide de tordre le dispositif initial en modifiant les produits d'emprunt proposés (en accord avec AAA) et en changeant le protocole d'enquête en cours d'expérimentation.

Un premier ajustement consiste à modifier les produits. Le protocole initial prévoyait une offre rurale similaire à l'offre proposée aux populations urbaines. Cette offre incluait pour l'essentiel du microcrédit solidaire auquel était adossé un délai de grâce de remboursement de deux mois dans le cas d'activités d'élevage. Il n'existe alors pas de produit spécifique pour les activités agricoles. Les autres produits sont des prêts individuels pour financer les micro-entreprises ou le logement mais les prêts individuels ne sont pas une priorité d'AAA dans ces zones. Alors que le document de projet initial mentionnait la nécessité de ne pas modifier les produits[1], les multiples changements effectués modifient de fait les services proposés (Naudet *et alii,* 2012) :

- des campagnes de sensibilisations hebdomadaires, bi-mensuelles ou mensuelles. Leur fréquence étant déterminée par le taux de participation ;
- alors que la procédure initiale prévoyait un quota minimal de femmes parmi les emprunteurs, ce quota n'est plus appliqué ;
- les agents bénéficient désormais d'une prime exceptionnelle de 10 dirhams par microcrédit déboursé et du remboursement de leurs frais de transports ;
- des campagnes de sensibilisation et de formation sont également prévues auprès des coordinateurs régionaux et des agents de terrain.

D'après nos interlocuteurs, la participation devient en effet une véritable « obsession » pour le personnel d'AAA également préoccupé par des questions d'ordre éthique concernant les ménages non servis.

Le second ajustement mis en œuvre par J-PAL pour pallier les conséquences statistiques de la faible participation (capacité à détecter des impacts) consiste à modifier le protocole de la RCT. Le modèle de prédiction de la probabilité de devenir client pour identifier les ménages les plus enclins à emprunter se base sur les résultats de l'enquête des sites de la première vague d'ouverture d'antennes rurales, soit sur quelques dizaines d'observations. Ce modèle s'avère cependant inopérant pour les vagues suivantes, notamment car les caractéristiques des zones retenues pour les différentes vagues évoluent et les variables ne permettent pas d'identifier les ménages qui présentent les plus grandes chances de prendre un microcrédit (AFD, 2009 ; DI_4). De même d'après l'AFD, la cohésion sociale à l'intérieur des villages n'a pas été prise en compte par l'équipe de chercheurs alors qu'elle constitue, selon les responsables de l'étude à l'AFD, un frein majeur dans la formation des groupes pour le microcrédit solidaire et donc de la faiblesse du *take-up* : « *Le diagnostic des agents est que le problème réside dans la constitution des groupes d'emprunteurs, car il y aurait un problème de solidarité dans ce douar*." (AFD, 2008 ; DI_3, p. 5).

[1] « *les agents ne doivent pas […] changer ou assouplir les règles d'obtention du microcrédit* » (J-PAL, 2006 ; DI_1, p. 26)

Le comité de pilotage de janvier 2009 revient à nouveau sur ce problème de participation, ainsi que d'autres difficultés. Le rapport mentionne une participation de l'ordre de 12% pour les vagues 2, 3 et 4 (contre 25% pour la vague 1), et une hétérogénéité encore plus forte (certains villages ont des taux d'emprunt de 80%) et « *cette hétérogénéité est pour l'instant mal expliquée par les caractéristiques des villages collectées* » (AFD, 2009 ; DI_4, p. 2). Le rapport indique également : « *L'utilisation d'une méthode d'échantillonnage permettant de sélectionner les ménages ayant les plus fortes chances d'emprunter a permis de pallier cette faible pénétration du micro-crédit, mais dans des proportions insuffisantes pour permettre de détecter un impact sur la consommation du programme inférieur à 26%. La faiblesse de cette correction est en partie liée à l'extrapolation des comportements de crédit des ménages de la vague 1 pour prédire les comportements de crédit des vagues suivantes* » (AFD, 2009 ; DI_4, p. 2).

Ce constat incite l'équipe à ajouter un nouveau sous-échantillon de ménages en *endline* (cf. partie 1). Ces ménages supplémentaires et leur répartition entre villages à forte, moyenne et faible pénétration devaient permettre de détecter un impact du microcrédit sur la consommation si celle-ci augmente de plus de 20%. Cette amélioration est toutefois conditionnée « *au maintien des efforts de promotion du micro-crédit par les agents d'Al Amana, pour les vagues 3 et 4* ». (ibid., p. 4). L'équipe de J-PAL envisage même l'abandon de plusieurs sites en raison des taux participation quasi nuls (DI_2). Or on peut supposer que les villages où les gens ne veulent pas s'endetter présentent des caractéristiques spécifiques. Comme indiqué en partie 1, la question des caractéristiques réelles de la population évaluée se pose donc (voir également Naudet, 2012 ; I_13 : p. 5).

Ces multiples amendements au protocole initial posent de sérieux problèmes. Outre qu'ils ne sont pas exposés de manière transparente et alors qu'ils auraient mérité une discussion scientifique, ils introduisent des biais dans un dispositif expérimental censé n'en avoir aucun. D'après l'AFD cela pose la question du « qui » et du « quoi » pour deux principales raisons. D'une part, du fait de la sélection des ménages qui ont une probabilité de devenir client selon un score *ad hoc* et inapte à estimer la propension à emprunter et, d'autre part, car les produits ont évolué tout au long de l'évaluation (Naudet *et alii*, 2012).

IIC- Des données de très mauvaise qualité, pointée du doigt par le bailleur

Comme prévu dans le document initial du projet, la collecte des données est sous-traitée au bureau d'étude *Team Maroc*, spécialisé dans l'ingénierie et dont la conduite d'enquêtes statistiques est loin d'être la spécialité[1]. Ce bureau est mandaté pour la passation des questionnaires et la saisie des données. Or, l'équipe AFD qui comprend des chercheurs expérimentés dans le domaine spécifique de la collecte de données avait pris la peine d'aller sur le terrain au mois d'avril 2008 pour observer le déroulement des enquêtes. Le compte-rendu rédigé à l'issue de sa mission de suivi laisse apparaître de sérieux dysfonctionnements (AFD, 2008 ; DI_3). Le rapport mentionne de nombreux problèmes de traduction des enquêtes car les enquêteurs ne maitrisent pas le berbère, langue pourtant parlée par

[1] Fondé en 1981, Team Maroc est devenu le premier groupe d'ingénierie privé indépendant du Maroc certifié ISO9001. A travers quatre domaines d'intervention – le bâtiment, les infrastructures, le conseil et les études sectorielles, «*Team Maroc, en complément à ses qualités techniques déjà largement reconnues, offre désormais a ses clients l'accès à l'expertise internationale du groupe Jacobs, notamment en terme de management de projets, et de sécurité sur les chantiers* » (http://archimedia.ma/a-la-une/communiques-de-presse/3435-team-maroc-appelle-a-renforcer-le-role-de-lingenierie-dans-le-developpement-national-a-loccasion-dune-table-ronde-a-rabat).

une grande partie de la population vivant dans les zones enquêtées. En effet, d'après le recensement général de 2014, entre 30 et 40% de la population marocaine parle le berbère et 16% de la population ne parle que le berbère (lequel comprend des dialectes distincts). Ce taux atteint 80% voire 100% dans certaines zones rurales reculées[1]. Ce problème aurait pu être anticipé par J-PAL et comme l'avait fait l'équipe de l'étude qualitative. Les enquêteurs ont donc largement recours à des traducteurs improvisés, induisant des problèmes de compréhension et des biais de réponse. Ainsi, le rapport de l'AFD soulève un cas pour lequel c'est l'autorité locale (le *Mokadem*) qui est amenée à jouer ce rôle sans précaution de la part des enquêteurs.

Le rapport souligne par ailleurs le problème de la multiplicité des répondants au sein des maisonnées et des familles élargies, qui semble se faire, là encore, de manière improvisée en fonction de la présence des uns et des autres et de leur aptitude à prendre la parole. Ces éléments peuvent fournir des éléments d'explication des décalages importants observés sur les informations recueillies au niveau de la composition des ménages entre la *baseline* et la *endline* que nous évoquons dans la première partie de cet article. Leur ampleur invite cependant à envisager l'hypothèse qu'au-delà de la variété d'interlocuteurs, il ne s'agisse tout simplement pas des mêmes ménages, comme nous le montrons dans Bédécarrats *et alii* (2019). L'absence d'adresse précise suppose des techniques précises d'identification (*tracking*) qui ont peut-être été négligées. Ou bien pressés par le temps (et insuffisamment supervisés), certains enquêteurs se sont contentés d'interroger les ménages disponibles au moment de leur passage.

A l'issue de cette mission, l'équipe AFD prend le soin de formuler de nombreuses pistes pour améliorer la qualité des données collectées. Ce rapport sera même suivi d'une lettre adressée à François Bourguignon, directeur de l'Ecole d'économie de Paris, datée du 19 mai 2008, dans laquelle l'AFD fait part de ses inquiétudes quant aux conséquences potentielles de ces failles sur les résultats de l'étude. Le courrier soulève également des problèmes à l'étape de saisie des données : les corrections, lorsqu'elles sont apportées, sont faites arbitrairement sans nécessairement recourir aux questionnaires[2]. Ce courrier fera l'objet d'une réponse de J-PAL le 19 juillet 2008, qui conteste la gravité des difficultés et atteste qu'elles ne remettent pas en question la validité interne de l'étude. Il n'en reste pas moins que les questionnaires déjà saisis au moment du Comité de pilotage de l'étude 29 janvier 2009, soit toutes les enquêtes de la *baseline* et une partie de celles de la *endline,* ont dû être envoyés à l'Insee afin d'être ressaisis. Cela marque de l'ampleur du désastre (DI_4).

Le rapport du comité de pilotage de janvier 2009 revient néanmoins sur la mauvaise qualité des données collectées (AFD, 2008 ; DI_4). Lors de ce comité, l'équipe de J-PAL attribue le problème à une insuffisance de moyens financiers et humains (formation, rémunération, personnel). Pourtant, les enquêteurs bénéficiaient d'une formation universitaire solide pour ce type de travail « bac+4 et +5 » (AFD, 2008 ; DI_3), mais ils n'ont visiblement pas été formés (ou pas suffisamment) aux techniques spécifiques d'enquêtes, et aux particularités de cette enquête en particulier, comme l'exige toute enquête. Par ailleurs, le budget sollicité par-J-PAL avait été intégralement alloué par l'AFD.

[1] Voir http://www.axl.cefan.ulaval.ca/afrique/maroc-1demo.htm, dernière mise à jour le 23 novembre 2018. Le berbère n'a été réhabilité comme langue officielle qu'en 2011 (Bouhjar, 2012).

[2] Le compte-rendu de mission d'avril 2008 évoquait amplement ce problème de modification arbitraire des informations lors de la saisie, Team Maroc justifiait ces changements en invoquant le « bon sens » (DI_3 : 4).

Les chercheurs soulignent à nouveau lors du Comité de pilotage de 2009 que « *La longueur des questionnaires, la faible motivation des enquêteurs et le manque de supervision peuvent affecter la qualité des données recueillies* » (AFD, 2009 ; DI_4). A la demande du chef de projet de J-PAL, il sera prévu pour le restant de l'enquête (mais celle-ci est déjà bien engagée puisque la *endline* a démarré) 1) de recruter des superviseurs supplémentaires (extérieurs à Team Maroc) pour vérifier la qualité des données, p. 2) d'octroyer des primes en fonction de la qualité des données, p. 3) d'augmenter le montant de rémunération par questionnaire ; p. 4) et dans chaque village de traitement d'organiser une « *visite jointe de l'agent local d'AAA et du superviseur pour s'assurer du bon appariement des bases de données* ».

En définitive, ce suivi au plus près du terrain et de la multiplicité d'acteurs engagés, qui ont tous leurs propres motivations et contraintes, donne un premier éclairage sur les défaillances listées en première partie : des enquêtés qui peinent à être convaincus de l'intérêt du microcrédit et n'en veulent pas (ce qui conduit à faire dérailler le protocole) ; des organisations de microcrédit concurrentes qu'il faut convaincre de ne pas pénétrer les zones d'enquêtes mais qui y sont en fait déjà, et qui s'en retireront, soit sous l'influence d'AAA (et de son directeur), soit du fait de la crise (ce qui conduit à changer l'objet d'étude, qui finalement étudie la substitution d'AAA à d'autres crédits formels) ; des agents de crédit « obsédés » par la participation (ce qui conduit à transformer les produits de microcrédit) ; un bureau d'étude non spécialisé et des enquêteurs visiblement peu formés et peu motivés, ne parlant pas nécessairement les langues locales (entachant la qualité des données) ; des membres des familles enquêtées qui interfèrent dans les questionnaires et des leaders locaux sollicités comme traducteurs de fortune (générateurs de biais de réponse). Ces points, qui auraient largement pu être évités, mettent surtout en lumière un défaut de compétence et de savoir-faire en matière d'enquête de la part de J-PAL et de ses prestataires.

III. Les coulisses de la fabrication et de la diffusion scientifique de la RCT-AAA : l'art de l'omission, de la persuasion et du contournement

La partie précédente a montré à quel point il était important d'explorer les rouages et coulisses multiples de la fabrique du savoir scientifique pour mieux comprendre*, in fine*, le degré de fiabilité des données collectées, depuis l'échantillonnage jusqu'aux estimations, en passant par la saisie. Cette exploration a également permis de lever l'un des trois paradoxes mentionnés en introduction (simplicité affichée de la méthode *versus* complexité du protocole réel) et d'esquisser quelques éléments factuels concernant le deuxième paradoxe, à savoir le contraste entre les conclusions des deux équipes impliquées (J-PAL *versus* AFD), dont l'une est finalement l'antithèse de l'autre. Cet exercice est également utile pour comprendre les conditions et modalités de dissémination des résultats de la RCT-AAA, et ainsi proposer des éléments d'interprétation du deuxième paradoxe, puis résoudre le troisième : comment justifier le succès d'une étude empirique qui contredit pourtant en bien des manières les exigences de base de la rigueur scientifique. Un détour par la fabrique du fait scientifique (c'est-à-dire le processus de transformation des matériaux empiriques en résultats scientifiques) est ici très utile pour compléter les éléments apportés dans la partie précédente : il met en évidence une stratégie particulièrement offensive menée sur l'ensemble du champ de la recherche et permettant aux auteurs de CDDP de faire table rase du passé, de s'affranchir de la critique et de contourner les règles de base de l'exercice scientifique.

La force d'un énoncé scientifique, c'est-à-dire sa capacité de persuasion, relève avant tout des conditions sociales, matérielles et politiques de sa production, des chaînes de traduction qu'il mobilise – comment cet énoncé transforme des observations en théorie–, et de leur inscription dans des réseaux plus ou moins influents où se mêlent de multiples acteurs (Akrich *et alii,* 2006). Cette capacité de persuasion est également le fruit de luttes historiques entre acteurs (et les écoles de pensée qu'ils représentent) pourvus de capitaux spécifiques, qui déploient des stratégies diverses pour asseoir ou renforcer leur crédibilité, non seulement dans le domaine de la production scientifique en tant que telle, mais dans l'ensemble des institutions qui gouvernent et régulent la pratique scientifique (Bourdieu, 1976). Cette lutte mobilise la démonstration scientifique, mais aussi des stratégies de positionnement et d'investissements, y compris matériels, ainsi que des procédures subtiles de formulation. Latour et Woolgar (1996), à partir de l'étude ethnographique d'un laboratoire américain de neuroendocrinologie, étudient le processus de traduction de données empiriques en « faits scientifiques » et le décrivent comme une suite « d'inscriptions littéraires » visant à persuader et convaincre. Il ne s'agit pas, pour les deux auteurs, de réfuter l'existence de faits, mais d'insister sur la dimension arbitraire, subjective et contingente de leur construction, qu'ils considèrent comme une tendance récurrente : une fois le produit final obtenu (c'est-à-dire le fait ou l'énoncé scientifique), l'ensemble des étapes intermédiaires et des procédures matérielles qui ont permis sa production sont reléguées au rang de la « technique » (Latour et Woolgar, 1996, p. 58). Des faits émergent – ici, les impacts du microcrédit – et deviennent dissociés des techniques matérielles (échantillonnage, enquêtes, codage, économétrie) qui leur ont pourtant donné naissance (ce que Bachelard (1965) nomme la « phénoménotechnique », qu'il considère comme une caractéristique de la science moderne).

Par le recours à l'assertivité, à l'art de la citation et du positionnement (s'appuyer sur des travaux existants, les réfuter ou les dénigrer), certaines opérations successives de traduction peuvent transformer une assertion qui pourrait sembler spéculative en une affirmation irréfutable. Cet art de l'inscription littéraire et de la persuasion est au cœur de la lutte entre chercheurs, laboratoires et écoles de pensée, de manière décuplée dans le champ des sciences sociales. Or cette lutte est foncièrement asymétrique et fonctionne de manière cumulative, à l'image du concept de capital de Bourdieu (1976). Un énoncé est d'autant plus crédible qu'il est porté par des chercheurs dont la crédibilité est déjà reconnue. Et un chercheur déjà crédible peut beaucoup plus facilement recourir à l'assertivité et au dénigrement (Latour et Woolgar, 1996, p. 163). Le processus est d'autant plus cumulatif que la recherche est un « marché » spécifique où les « producteurs » sont également les « consommateurs » : la crédibilité scientifique ne provient que des pairs (Bourdieu 1976).

Appliquer ce type d'analyse à la RCT-AAA nous éclaire sur les trois paradoxes que l'on cherche à comprendre. Nous suggérons que le décalage entre l'approximation voire la désinvolture de l'application de la méthode (patente dans le cas étudié ici) et sa force de persuasion découlent d'une stratégie globale de la part des auteurs de CDDP (et plus globalement de J-PAL) visant à s'arroger le monopole de la scientificité en matière d'étude d'impact (voir également Jatteau, 2016 et 2018). L'équipe de J-PAL étant à la fois initiateur, opérateur, producteur et diffuseur de la RCT-AAA, cette stratégie est présente à toutes les étapes du processus et se caractérise à la fois par l'omission, l'assertivité, et enfin le contournement des règles de base de l'exercice scientifique.

III-A. L'art de l'omission et de la persuasion

Le défaut évident de rigueur dans la collecte, décrit dans la partie précédente, illustre déjà une mésestimation des compétences nécessaires, un désintérêt pour la fiabilité et la qualité des données et plus globalement une absence de « culture de la donnée ». La transformation de données douteuses en faits scientifiques s'appuie ensuite sur une série d'opérations (d'inscriptions littéraires) caractérisées à la fois par l'assertivité et le déni de l'existant : l'affirmation du caractère irréprochable de la méthode (en dépit de critiques répétées et rendues publiques par l'AFD)[1] ; l'affirmation de la dimension inédite des résultats (qui implique, là encore, un refus de toute forme de controverse et un mépris de l'existant, et alors même que les résultats, s'ils étaient valides, apportent peu à l'état des connaissances) ; et enfin des traductions successives qui relèvent de l'extrapolation plus que de l'interprétation rigoureuse.

S'affranchir d'une « culture de la donnée »

Les tenants des RCT en économie du développement les ont importées du champ médical en faisant abstraction des discussions critiques, des conditions de mises en œuvre et remises en cause dont elles ont fait l'objet en santé publique (Eble *et alii*, 2014 ; Spears et *alii* à paraître ; Garchitorena et *alii,* à paraître ; Kraus, 2018). Ils font également abstraction des débats propres à la collecte de données. Dans la plupart des protocoles de recherche empirique de nature quantitative, il existe une division du travail entre les collecteurs de données et les analystes : les premiers sont des statisticiens, les seconds des économistes (économètres ou spécialistes thématiciens), et à quelques exceptions près (par exemple Deaton 1997 ou Grosh et Glewwe, 2000) rares sont ceux qui tiennent les deux bouts du spectre. Il s'agit là de métiers à part entière, supposant des compétences et des apprentissages distincts. Les uns se chargent de la précision de la mesure, les autres de sa pertinence, de son analyse et des liens entre données, les deux étant indispensables à la production finale de résultats raisonnables, même si le prestige académique des premiers est moindre que celui des seconds (Desrosières, 2013). Ici, les erreurs multiples de collecte (dont certaines ont été pointées de manière réitérée, par l'équipe de l'AFD, qui comptait dans ses rangs des spécialistes des enquêtes ; cf. supra) et de saisie témoignent d'un manque évident d'expérience et de connaissance, comme si les compétences purement techniques en amont (économétrie : traitement des biais et sélection et identification d'un contrefactuel) libéraient les chercheurs de tout le savoir-faire nécessaire à la collecte de données statistiques de qualité en aval.

La déconnexion entre les chercheurs et le terrain en est une autre illustration. Cette déconnexion résulte d'une organisation hiérarchique au sein des laboratoires de recherche dont découle une division stricte du travail entre chefs de projets, doctorants et personnel de terrain (superviseurs, enquêteurs). Ces derniers se voient finalement attribuer une responsabilité considérable, pour laquelle ils ne sont pas formés de manière adéquate (cf. supra ; Jatteau, 2016). Cette division du travail est un mode d'organisation que l'on retrouve fréquemment en sciences expérimentales de la matière et de la vie, mais qui n'empêche pas les chefs de projets de rester régulièrement au contact

[1] En fait, l'équipe de l'AFD a fait état de ses interrogations critiques à l'équipe de J-PAL, lors de leurs interactions bilatérales, mais s'est gardée d'en faire mention dans les articles publiés sur le sujet, de peur de s'attaquer à un « trop gros morceau ». Ce « double langage » est une illustration directe de l'effet d'imposition exercé par les institutions et les chercheurs prestigieux du champ, dans la confrontation asymétrique décrite plus haut (et p.25).

de la chaîne de production de données, y compris pour des expérimentations *in vivo*. Et ils sont tenus de suivre des protocoles précis de validation de la rigueur des expériences menées. Cela n'est pas le cas ici. Cette déconnexion va ensuite être aggravée par l'expansion exceptionnellement rapide de J-PAL (Ravallion, 2018), sur laquelle nous revenons plus loin.

Enfin, toute enquête quantitative suppose une étape préalable exploratoire, sous forme d'enquêtes pilotes, éventuellement d'analyses qualitatives, permettant d'élaborer ensuite un protocole et des questionnaires adaptés aux réalités locales. Visiblement, cette étape a été négligée[1] et ceci explique certainement une partie des défaillances décrites en partie 1.

Concernant la mauvaise qualité des données et les décalages multiples entre *baseline* et *endline*, et outre les faiblesses dans la conduite des entretiens (cf. supra), on peut émettre l'hypothèse que l'équipe de J-PAL a sous-estimé les complexités locales, ceci ayant conduit à des approximations et des incohérences entre vagues d'enquêtes (et entre ménages). Par exemple, la notion de « ménage » renvoie à des « constellations familiales » polymorphes et non à une réalité figée. Les écarts considérables observés entre les deux vagues proviennent, soit d'imprécisions dans les questions posées, soit d'erreurs dans le choix des ménages enquêtés lors de la seconde vague. Des difficultés identiques se posent pour la notion d'actifs. Dans certaines régions, notamment à dominante berbère, les décisions relatives à l'élevage ne sont pas individuelles mais collectives, précisément pour tenir compte des contraintes liées au pâturage et au gardiennage des animaux. Dans ces mêmes zones, les parcelles, y compris celles de propriété individuelle (*Melk*), sont rarement redistribuées afin d'éviter le morcellement, mais les dépenses et la production sont fréquemment partagées entre frères, y compris lorsqu'ils ont migré. De telles spécificités brouillent les unités d'analyse (ménage, actifs, etc.). Si l'analyse statistique ne peut évidemment pas rendre compte de toute la complexité des situations, elle doit néanmoins harmoniser la manière dont ces singularités liées au contexte sont appréhendées dans l'enquête afin d'assurer la cohérence des informations collectées.

L'équipe de J-PAL a choisi d'utiliser le fichier clientèle de l'IMF plutôt que l'enquête pour identifier le groupe de traitement, alors que ces données n'étaient pas fiables, notamment du fait de pratiques courantes de prête-noms (Bédécarrats *et alii*, 2019). Là encore, une enquête exploratoire aurait pu prendre la peine de vérifier la fiabilité de ces données. Concernant les erreurs de sondage, on a vu qu'elles découlaient de distorsions multiples du protocole suite à la faiblesse de la participation. Là encore, des enquêtes exploratoires auraient certainement pu anticiper cette difficulté majeure.

Affirmer l'irréprochabilité de la méthode et ignorer la critique

Tandis que l'équipe de J-PAL s'appuie sur cette étude pour construire un discours à vocation universaliste sur l'impact du microcrédit, l'AFD s'en saisit *a contrario* pour se forger une expertise sur la méthode et conclure que les RCT sont dans bien des cas inadaptées pour mesurer l'impact de projets de développement. Le hiatus entre le diagnostic des deux équipes est patent : c'est le deuxième paradoxe souligné en introduction. Mais l'équipe de J-PAL passe outre.

Dès 2009, alors que la *endline* est toujours en cours, l'AFD commence à communiquer sur son expérience en matière de RCT (Delarue, 2009 ; I_9). Son diagnostic, issu du cas Al Amana et conforté par une autre étude menée parallèlement au Cambodge, est clair : il souligne la grande difficulté de

[1] Le document de projet mentionne des enquêtes qualitatives mais reste très allusif (J-PAL, 2006 ; DI_1).

la méthode à évaluer l'impact, compte tenu des multiples entorses au protocole que l'équipe AFD a en partie repérées (problème de représentativité, modification du produit) et des contraintes de temporalité qui obligent à se focaliser sur le court terme. Les contraintes de la méthode RCT engendrent des perturbations opérationnelles diverses, incompatibles en définitive avec l'évaluation d'impact d'un projet spécifique, dans le cadre des activités courantes d'un bailleur de fonds. Indéniablement déçus des résultats de l'étude, dont l'équipe AFD doute fortement de la validité[1], elle essaie néanmoins de mettre à profit cette expérience pour en tirer des leçons sur les limites de la méthode.

Ce diagnostic sera ensuite partagé lors de plusieurs événements internationaux (DIME[2], NONIE[3], CGAP[4]) et à plusieurs reprises (12 interventions publiques, nationales et internationales, recensées entre 2009 et 2013). Il fera l'objet d'un article scientifique publié relativement rapidement (2012), avec un effort de diffusion à la fois en langue française, dans la *Revue d'Economie du Développement*, l'une des meilleures revues francophones du champ (Naudet *et alii*, 2012) et une version anglaise dans une revue précisément spécialisée sur les questions d'impact et d'efficacité des politiques de développement (Bernard *et alii*, 2012). Dans ces deux articles, les commanditaires de l'étude AFD considèrent qu'en définitive, les RCT ne sont adaptées qu'à des projets circonscrits, que les auteurs qualifient de « projets tunnels », caractérisés par des impacts de court terme, des inputs et outputs clairement identifiés, facilement mesurables, des liens de causalité unidirectionnels (A cause B), linéaires et enfin non soumis à des risques de faible participation de la part des populations visées. Bien qu'il s'agisse d'interventions relativement simples, le microcrédit n'en fait pas partie. Dans une présentation à la conférence NONIE de 2012, Jean-David Naudet, alors chef du département Evaluation de l'AFD, concluait sans équivoque : « *Les EI- évaluations d'impact - ne seront pas l'outil par lequel l'AFD sera redevable de son action* » (Naudet, 2012 ; I_13, slide 11). L'étude AAA alimente aussi un rapport publié en 2015 par l'AFD et dédié à l'évaluation d'impact (Pamiès-Sumner, 2015). Il resitue les enseignements AFD dans des débats plus larges parmi la communauté des évaluateurs, au sein de laquelle un consensus semble émerger en faveur du pluralisme méthodologique, et la nécessité de considérer les RCT non comme plus comme le "*gold standard*" mais comme un "*good standard*" (Pamiès-Sumner, 2015).

A défaut d'obtenir des résultats fiables, cette expérience aura donc permis à l'AFD de se forger une expertise sur le sujet, à la fois en interne – l'AFD arrêtera de financer des RCT, son comité des évaluations entérinera l'idée de leur privilégier d'autres méthodes, mixtes si possible – mais aussi de

---

[1] Un de responsable de l'étude en poste à l'époque résume les choses de la manière suivante « *on mesure quelque chose mais on ne sait pas ce que c'est, sur qui c'est et quand c'est* » (Echange d'emails, 18 février 2018). Sur la mauvaise qualité des données et leurs usages, dont l'AFD avait bien conscience, notre interlocuteur mentionne également : « *[…] on peut parler des papiers finaux où les résultats commentés relèvent sans doute bien plus du hasard que d'autre chose et où on peut s'amuser (ou se désoler) à voir […] essayer d'inventer des histoires de développement pas crédibles pour essayer de coller à leurs résultats (effet sur les salaires et pas sur les revenus ou quelque chose dans ce genre)* ».

[2] DIME (Development Impact Evaluation) est le groupe de recherche de la Banque mondiale en charge des évaluations d'impact, de projets menés par la Banque mondiale ou par d'autres institutions.

[3] NONIE (*Network of Networks for Impact Evaluation*) regroupe l'ensemble des réseaux de promotion des études d'impact.

[4] CGAP (Consultative Group to Assist the Poor) est un réseau (informel mais néanmoins hébergé par la Banque mondiale) regroupant certains des praticiens et bailleurs les plus en vue du secteur de la microfinance, et qui joue en rôle central en matière de diffusion de bonnes pratiques.

contribuer en externe à certains débats internationaux, aussi bien dans la sphère de l'évaluation que celle des praticiens du microcrédit *via* le CGAP. Bien que l'équipe AFD ait eu parfaitement conscience de la qualité douteuse du protocole et des données - tout au long de l'étude ils n'ont eu cesse, nous l'avons vu, d'inciter les chercheurs de J-PAL à rectifier le tir[1] – ils choisissent délibérément de ne pas communiquer sur ce point[2]. Ce choix vise probablement à ne pas heurter de front l'un des acteurs les plus puissants à l'échelle mondiale dans le champ de l'évaluation et de la recherche afférente (J-PAL), ainsi que les chercheurs de haut niveau impliqués dans l'étude, et éventuellement à donner plus de poids à leur propre critique, qui est d'un autre ordre.

Le numéro spécial d'AEJ: AE qui publie les résultats et vise une toute autre audience (des économistes universitaires), ne fait nullement mention de ce type de débats. Si L'AFD parvient en partie à tirer son épingle du jeu, l'asymétrie des positions ne fait toutefois aucun doute : la version anglaise de l'article écrit par l'équipe AFD n'est à ce jour (février 2019) citée que 15 fois, et quasiment exclusivement par les tenants d'une critique forte. La version française a été citée 17 fois.

Non seulement l'article publié par CDDP n'y fait nullement mention, mais il fait également fi de toutes les entorses au protocole initial. Toutes les pratiques scientifiques empiriques « bricolent », au sens où les contraintes de terrains supposent adaptation, compromis, approximations et imperfections. Les études randomisées ont toutefois deux spécificités. La prétendue simplicité et pureté de la méthode (simple comparaison de moyennes) sont présentées comme des arguments de supériorité à l'égard de méthodes plus classiques, comme les modèles macroéconomiques ou la microéconométrie sur données d'observations ou quasi-expérimentales. Or cet argument de simplicité contraste avec l'extraordinaire complexité de la méthode finale décrite dans les parties précédentes, liée à la fois aux exigences de la randomisation et à la multiplicité des acteurs impliqués (le troisième de nos paradoxes mentionnés en introduction). C'est précisément cette complexité qui rend les RCT particulièrement sujettes au bricolage.

Dans l'étude dont il est question ici, les écarts observés entre protocole initial et réel aboutissent à des données et un échantillonnage de qualité discutable : il ne s'agit plus de bricolage mais de négligence et/ou d'incompétence (cf. supra), voire de désinvolture guidée par l'objectif de publier à tout prix (cf. infra). D'autre part, et alors que la réflexivité est partie intégrante de toute démarche scientifique (depuis le journal de laboratoire du chimiste jusqu'aux carnets de terrain du socio-anthropologue), les randomisateurs du développement en font très rarement état, mise à part dans des blogs spécialisés où ils s'échangent leurs tours de main (Bédécarrats *et alii*, 2018). Ici en l'occurrence, et en dépit des alertes répétées (et publiques) de l'équipe de suivi de l'AFD (leur article est publié, rappelons-le, en 2012, soit 3 ans auparavant), les chercheurs du J-PAL y restent sourds. Dans leur propre article, l'équipe de J-PAL passe entièrement sous silence les entorses multiples

[1] Outre les éléments mentionnés plus haut, de nombreuses discussions informelles avec les différents membres de l'équipe AFD attestent de leur désarroi à l'égard du manque de rigueur du protocole. « *J'ai honte de cette étude »,* nous confiait l'un d'entre eux en 2010, alors que nous préparions notre propre terrain.

[2] Dans cette même communication de 2012 donnée à la conférence NONIE, la présentation de Naudet indique « *Une leçon à partager : les EIs financées par l'AFD ont été faites* dans les règles de l'art*, mais n'ont pas permis d'évaluer les bénéfices à attendre des interventions concernées* » (Naudet, 2012 ; I_13, slide 3, nous soulignons).

faites au protocole[1]. Elle le fait probablement de manière délibérée pour ne pas entacher la rigueur affichée de l'étude RCT-AAA, et elle peut se le permettre : son capital déjà acquis l'autorise à s'affranchir des critiques (Bourdieu, 1976), même si celles-ci ont été rendues publiques.

Revendiquer l'inédit et omettre l'existant

Faire table rase des recherches antérieures participe également de l'offensive. Comme nous l'avons montré par ailleurs (Bédécarrats *et alii*, 2018), les résultats sont souvent présentés comme des « découvertes » inédites, alors qu'il ne s'agit fréquemment que de la reprise de conclusions obtenues par des études antérieures. Le subterfuge est d'autant plus facile à faire admettre que les publications tirées de méthodes non expérimentales ne sont presque jamais citées (Labrousse, 2010 ; Nubukpo, 2012).

L'article CDDP contient 17 références : 3 proviennent des auteurs (18%), 9 des membres de J-PAL (53%), et les références restantes relèvent pour moitié de technique statistique (4), et de la microéconomie des ménages (1 purement théorique et 3 qui s'appuient sur des données empiriques). Les connaissances accumulées dans le passé sur le thème du microcrédit et issues de méthodes non randomisées s'apparentent donc à un champ vierge. Comparés à l'aune de l'existant (et indépendamment des problèmes de validité), les résultats sont pourtant moins novateurs qu'ils n'y paraissent.

Prenons quelques exemples. Un résultat majeur concerne les effets limités et hétérogènes du microcrédit en matière de micro-entrepreneuriat. Or ceux-ci sont discutés depuis deux décennies. Critiqués pour leur focalisation sur des effets moyens, les randomisateurs ont tenté d'appréhender l'hétérogénéité et admettent son existence, mais sans être en mesure d'en définir les modalités précises ni de l'expliquer. La question posée de manière récurrente aux RCT reste donc entière : en matière de politiques de développement, quel sens peut-on donner à un impact moyen, et celui-ci a-t-il réellement un sens ? Sur ce point, comme sur les effets limités en matière d'entrepreneuriat, la littérature disponible apporte des éclairages bien plus convaincants. Si la quantité n'est évidemment pas un gage de qualité, on peut néanmoins remarquer qu'en 2011, 150 études d'impact non expérimentales (contre six RCT) avaient déjà été recensées (Bédécarrats, 2013). Détaillons quelque unes d'entre elles. L'étude pionnière et comparative de Hulme et Mosley (1996) pointait les limites du microcrédit sur l'entrepreneuriat, tout en interrogeant en détail la diversité des profils individuels et des contextes susceptibles d'expliquer l'hétérogénéité des impacts. Morduch (1999), dans un article également pionnier basé sur une revue de la littérature, concluait lui aussi à des effets limités, susceptibles au mieux de compléter des revenus existants (et non de modifier radicalement des portefeuilles d'activités), et largement dépendants de contextes plus ou moins propices à l'entrepreneuriat. D'autres travaux ultérieurs sont venus confirmer ces tendances générales (voir par exemple Armendáriz and Morduch, 2010 ; Bouquet *et alii*, 2007 ; Copestake *et alii*, 2001 ; Roodman 2011). Certes, les méthodes randomisées ont pour objet de lever l'épineuse question des biais de sélection (la fameuse attribution causale), dont l'importance est notoire et a été démontrée dans le cas du microcrédit (voir par exemple Coleman, 1999). Mais les exigences de la randomisation suscitent d'autres biais qui sont tout aussi problématiques, comme l'indique le point suivant.

[1] On remarque que les auteurs précisent en note: "*The draft was not reviewed by the AFD, IGC, or J-PAL before submission and only represents the views of the authors*" (CDDP, note *).

Revendiquer l'universalité et omettre les contextes

Compte tenu des multiples amendements au protocole d'échantillonnage, le profil de la population ciblée, nous l'avons déjà vu, est particulièrement obscur (voir également Cull et Morduch, 2017). Contextualiser reste donc incontournable pour expliciter le type de population étudié (Pritchett et Sandefur, 2015). Contextualiser est également incontournable pour donner vie aux données collectées, puisque les données « ne sont jamais données » (Desrosières, 2013). Notre réplication nous a déjà amenés à proposer un énoncé radicalement différent : l'étude ne compare pas clients et non clients, mais en fait la substitution d'Al Amana à d'autres sources de crédit déjà disponibles (cf. supra et Bédécarrats *et alii*, 2019). L'étude qualitative menée par deux d'entre nous alors que se déroulait la *endline* permet d'affiner cet énoncé. Cette étude portait, entre autres, sur les usages du microcrédit (Guérin *et alii*, 2010). Contrairement aux déclarations des clients (reprises par l'équipe de J-PAL)[1], nous observions un usage massif (de 60 à 80% selon les contextes) pour de la consommation courante, des biens durables et l'amélioration de l'habitat. L'usage limité pour l'entrepreneuriat non agricole (moins de 10%) et l'élevage (10 à 30%) s'expliquait de deux manières : une absence de débouchés pour le premier (sauf en zones péri-urbaines, et dans des proportions limitées) et de faibles possibilités d'expansion pour le second, du fait de contraintes fortes de main-d'œuvre et de pâturage. Ce constat n'était toutefois pas incompatible avec des impacts positifs en termes d'équilibre général (mais que notre protocole ne permettait pas de mesurer) : l'habitat, grand consommateur de matière première et de main-d'œuvre locale, est susceptible d'avoir de forts effets d'entrainement (contrairement au petit commerce de détail dont les effets de substitution sont bien connus). Nos conclusions étaient donc sensiblement différentes de celles de J-PAL, y compris dans l'analyse théorique des processus à l'œuvre, tout en ayant leurs propres limites, inhérentes aux analyses qualitatives[2]. Conscientes de cela, nous avions sollicité à plusieurs reprises l'équipe de J-PAL pour dialoguer et confronter nos méthodes et résultats. Illustration supplémentaire de leur rejet à l'égard d'autres modes d'administration de la preuve, nos tentatives sont restées lettre morte. Notons par ailleurs que l'usage fréquent des microcrédits pour des activités ne générant pas de revenu direct est désormais largement reconnu (voir par exemple Cull et Morduch, 2017). Or ce constat modifie sensiblement les chaînes de causalité sous-jacentes aux processus d'impact (et donc la théorie proposée par les randomisateurs, aussi bien dans l'article CDDP que dans l'introduction générale d'AEJ: AE).

De traductions en traductions : l'arbitraire des interprétations

---

[1] *"When applying for microcredit, most of clients (68 percent) declared to be planning to use the loan in animal husbandry activities, mainly cattle and sheep rising, 26.4 percent in trade-related businesses, and the remaining 5.5 percent in other non-agricultural businesses, such as services and handicraft. It is not surprising that no client declared an intent to allocate loans to other agricultural activities (crops and fruit trees), as Al Amana did not lend for such activitie*s » (CDDP, p. 134).

[2] Certains argueront que la mesure des têtes de bétail par ménage et sur un échantillon de grande ampleur est plus fiable que nos données qualitatives, tirées d'entretiens détaillés avec des clients, des non-clients, des leaders locaux et des agents de crédit. Mais les données déclaratives ont aussi leurs limites, *a fortiori* lorsqu'elles sont liées à des projets ou politiques de développement. Des clients ayant annoncé l'achat de bétail pour être éligible au crédit peuvent facilement être incités à surestimer leur nombre face à la crainte d'éventuelles représailles.

Un dernier point mérite qu'on s'y attarde, à savoir celui lié à la notion de « choix ». Les résultats relatifs à l'investissement entrepreneurial et au retrait de l'emploi salarié amènent les auteurs de la RCT-AAA à les commenter de la manière suivante : *« […] certains ménages choisissent de profiter du microcrédit pour changer, de façon assez significative, la façon dont leurs vies sont organisées. Mais même ces emprunteurs ne semblent pas choisir le microcrédit comme moyen d'augmenter leur niveau de vie, du moins à relativement court terme* » (CDDP, 2015, p. 127 ; notre traduction). Et c'est probablement ce résultat qui leur permet d'affirmer que le microcrédit est un « *instrument financier puissant pour les pauvres* » (CDDP, p. 148), alors que les résultats strictement économiques, nous l'avons vu, sont limités. L'usage répété du verbe « choisir » laisse à voir des entrepreneurs, hommes ou femmes, maîtres de leur destin, libres d'arbitrer entre plusieurs options et libres d'améliorer leur niveau de vie. Quelle est la signification du verbe « choisir » pour des familles montagnardes du Haut-Atlas, confinées à une agriculture de subsistance, *a fortiori* lorsque la notion de propriété individuelle a peu de sens, et qui vivent aux rythmes de la migration masculine et des aléas de la vie urbaine ? Admettons qu'elles se soient engagées dans une nouvelle activité, par exemple l'élevage, qui reste l'exemple le plus plausible dans le cas étudié. Que cette activité ne se traduise pas en termes d'amélioration du niveau de vie peut difficilement résulter d'un choix délibéré. Il peut aussi s'agir de maladies ou décès, *a fortiori* pour des personnes dépourvues de cheptel jusque-là, et donc probablement dépourvues d'expérience (tous les ruraux ne sont pas des éleveurs nés !) ; il peut s'agir, comme déjà indiqué, de difficultés d'accès à du pâturage ou du fourrage (impliquant des coûts supplémentaires ou une sous-nutrition, source de maladie), ou d'exigences de main-d'œuvre obligeant à renoncer à d'autres activités.

La sémantique a ici toute son importance, car les éditeurs du numéro spécial se saisissent de ce résultat pour conclure que la « liberté de choix » est en définitive un impact central du microcrédit, permettant de nuancer en quelque sorte les effets mitigés en matière de pauvreté. Or faute de détails contextuels, recourir au terme « choisir », puis interpréter cet élargissement des choix en « liberté », est en fait le fruit de traductions successives (traduction au sens de la sociologie des sciences, cf. Akrich *et alii*, 2006) qui relève plus de l'extrapolation que d'interprétations rigoureuses. Ce résultat positif en termes de liberté sera pourtant repris ensuite comme l'un des résultats principaux dans le *Policy Bulletin* de J-PAL et IPA (2015), visant à disséminer largement les résultats de J-PAL en matière d'étude d'impacts du microcrédit. Si CDDP se permettent cette traduction, c'est qu'ils mobilisent en fait une hypothèse implicite, qu'il semble inutile de préciser et de justifier tellement elle va de soi : la création d'activité ou d'entreprises est, en soi, synonyme de « liberté ».

III-B. Contourner certaines règles de la déontologie scientifique

Au déni de l'existant se rajoute l'évitement de certaines règles de base de la pratique scientifique. Si cet écueil semble grandissant dans l'ensemble de la communauté scientifique et n'est pas spécifique à J-PAL (Heckman et Moktan 2018), il est ici particulièrement manifeste.

Contourner l'évaluation par les pairs

Dans le monde de la recherche, la validation des connaissances repose sur le principe d'évaluation « par les pairs », qui désigne l'activité collective des chercheurs qui jugent de façon critique et anonyme les travaux des autres chercheurs de leur communauté. Ce principe est supposé garantir que seules des études de qualité rigoureuse soient publiées et ainsi reconnues. Diverses règles éthiques doivent toutefois d'être respectées, à commencer par la gestion des conflits d'intérêts entre

auteurs et membres des comités de rédaction des revues. Le favoritisme éditorial est un processus reconnu et démontré, notamment dans la communauté des économistes (Fourcade *et alii*, 2015). Il s'appuie habituellement sur des liens d'interconnaissance entre éditeurs et auteurs (formation et universités communes, co-publications, directions de thèse ; voir par exemple Colussi, 2018). Ici en revanche, on assiste à une véritable entorse aux règles de base en matière de conflits d'intérêt. L'article a été publié dans une revue fondée par Esther Duflo - *American Economic Journal: Applied Economics.* Elle en était l'éditrice au moment de la publication. Elle ne figure pas parmi les éditeurs du numéro spécial mais elle est co-auteure de deux articles. Deux des trois éditeurs du numéro spécial – Abhijit Banerjee et Dean Karlan – sont membres du comité éditorial et co-auteurs d'un article. Enfin près de la moitié des auteurs de l'ensemble des articles du numéro (soit 11 sur 25) sont également membres de J-PAL, et quatre autres y sont chercheurs associés ou doctorants. On s'éloigne ainsi singulièrement des principes de contrôle par les pairs qui doivent régir la publication scientifique. On est en droit de supposer que cette proximité étroite a relâché les exigences habituelles des procédures d'évaluation, ce qui expliquerait que l'article ait été publié en dépit de faiblesses multiples.

Croître et publier : une course en avant incompatible avec la rigueur scientifique ?

A ces défaillances diverses se rajoute, nous semble-t-il, une course en avant du laboratoire J-PAL (et de la communauté scientifique dans son ensemble) qui semble difficilement compatible avec la rigueur scientifique. Nous avons mentionné plus haut la division du travail entre chefs de projets et travailleurs de terrain. Cette division des tâches se combine avec un fonctionnement très hiérarchisé et centralisé du laboratoire qui garantit une conformité de pensée et de méthode et facilite l'expansion (Jatteau, 2018). J-PAL s'est en effet développé de manière très rapide par rapport à ce qui est habituellement observé dans le champ scientifique (Ravallion, 2018), et qui relève plutôt du monde de l'entreprise, accumulant en un temps record de nombreuses expérimentations (on en comptait 946 en février 2019, terminées et en cours). Cette course en avant vise certainement à se positionner sur le marché de l'évaluation d'impact mais aussi à accumuler des résultats permettant de pallier le défaut de validité externe qui est reproché aux randomistes de manière récurrente (Bédécarrats *et alii,* 2018). Or cette double caractéristique – centralisation/croissance rapide – implique qu'une poignée de chercheurs dirige un nombre considérable de RCT, ce qui interroge leur capacité effective d'implication (et renforce la déconnexion avec le terrain indiquée plus haut). Au moment de la rédaction de cet article (février 2019), Esther Duflo comptait 64 études randomisées à son actif, soit plus de 4 nouvelles études par an). On peut se demander quelle est son implication réelle dans chaque étude dont elle signe les résultats. Pour Bruno Crépon, cette étude est la première qu'il mène et son travail a visiblement été jugé satisfaisant puisqu'il n'a cessé d'en mener depuis (il en compte aujourd'hui 20) et est devenu un des nœuds centraux du réseau des randomistes sur le marché du travail (Jatteau, 2016, p.323). Concernant la RCT-AAA, le risque de déconnexion avec le terrain était probablement moindre qu'aujourd'hui puisque J-PAL en était encore à ses débuts (l'étude démarre 3 ans après la création de J-PAL). La pression à la publication, en revanche, est certainement présente. Les résultats de chaque RCT sont appréciés en termes de « *publishable unit* » et ne pas publier au moins un article par RCT est difficilement envisageable compte tenu des investissements consentis (Jatteau, 2016, pp. 423-432). Les auteurs de l'article ont-ils simplement été négligents ou ont-ils sciemment ignoré les principes déontologiques de base afin de publier, coûte que coûte ?

Comment expliquer l'affirmation d'une zone de contrôle « vierge », alors que nos propres analyses, comme leurs propres tableaux, montrent le contraire ? Comment expliquer le recodage étrange des sources de crédit, variable centrale, et qui permet aux auteurs d'affirmer, précisément, que la zone de contrôle était vierge ? Comment expliquer le choix sélectif des variables retenues pour attester de l'homogénéité entre population traitée et population de contrôle, alors que d'autres variables, pourtant centrales par rapport à la question de recherche posée (crédits formels, *assets*), font état de déséquilibres mais ont été omises de la présentation des résultats ? Comment expliquer l'inconsistance de la procédure de *trimming* ? Ce faisceau d'indices incite à penser que la pression à la publication a poussé les chercheurs à opérer une série de choix, contestables sur le plan de la rigueur scientifique, mais leur permettant d'affirmer que leur protocole était irréprochable et d'aboutir à des différences significatives entre leurs deux populations. Ces deux conditions sont essentielles pour prétendre publier. Si la première se justifie évidemment scientifiquement, il n'en est rien pour la seconde (l'absence d'impact est tout aussi instructive), or cette exigence implicite biaise la production de connaissances (Ravallion, 2018).

Ces pratiques ne sont bien sûr pas propres à J-PAL. On sait à quel point le diktat du « *publish or perish* » formate désormais le monde académique et engendre des travers multiples : saucissonnage, plagiat, duplication, signatures multiples ou encore fraudes. Les dérives observées ne font qu'illustrer les limites d'un système académique qui n'a plus les moyens d'apprécier la *qualité* des recherches menées (Heckman et Moktan, 2018). Alors que la plupart des chercheurs ont désespérément besoin de publier pour maintenir ou acquérir une place ou la maintenir dans le champ académique, d'autres – et c'est visiblement le cas de l'équipe de J-PAL – contournent les règles de déontologie pour exercer et renforcer une position de monopole.

Conclusion

Cet article avait pour but de décrire la chaîne de production d'un fait scientifique, depuis l'échantillonnage jusqu'à la publication et la dissémination des résultats, en passant par la collecte de données, la saisie et le recodage, les estimations et les interprétations, en s'appuyant sur un matériau inédit (une réplication des résultats de CDDP, une étude qualitative d'envergure menée en parallèle de l'expérimentation, l'exploitation d'un ensemble quasiment exhaustif des documents internes entre l'AFD et J-PAL, et des entretiens menés avec les acteurs). Il avait également pour but d'éclairer trois paradoxes : une médiatisation forte des résultats de la RCT-AAA dont la validité, tant interne qu'externe, est pourtant très problématique ; des résultats antinomiques entre les deux équipes de recherche impliquées : J-PAL s'appuie sur cette étude pour en tirer des conclusions universelles, tandis que l'AFD décide de ne plus retenir les RCT comme méthode privilégiée d'évaluation d'impact ; un protocole d'enquête extraordinairement complexe alors que la simplicité de la méthode est précisément l'argument phare de la supériorité des RCT.

Nos analyses révèlent, en partie, des processus classiquement observés en sociologie des sciences. Bricolage, traduction, positionnement stratégique (*via* les références à la littérature existante ou au contraire leur omission), concurrence et asymétrie des positions sont le lot commun de la production scientifique. Les pressions croissantes à la publication génèrent de nombreuses dérives. Contourner les règles d'évaluation par les pairs et s'affranchir des conflits d'intérêts n'est pas non plus exceptionnel, nous l'avons également évoqué. Plus largement, il a été montré que les économistes, de manière générale, éprouvent un sentiment de supériorité sur les autres sciences sociales, issu, entre autres, de leur plus forte proximité avec les sciences expérimentales, les entreprises et le

pouvoir politique ; ce sentiment de supériorité les conforte dans leur isolement épistémologique et dans leur revendication à « trouver des solutions » (Fourcade *et alii*, 2015).

Si les pratiques de l'équipe de J-PAL confirment ces différentes observations, elles se distinguent par ce que l'on pourrait qualifier d'*hubris*, combinant arrogance, sentiment d'impunité, et démesure (la course en avant évoqué précédemment). Or l'*hubri*s les amène à sortir du code de déontologie propre à la recherche scientifique. Les bricolages sont légion, mais à condition de faire de preuve de réflexivité. La traduction est inévitable, mais à condition d'en préciser les modalités et de les justifier. La concurrence rythme la vie scientifique, mais à condition qu'elle ne se transforme pas en monopole, et que ses dérives soient combattues. Or c'est précisément ce qui se joue ici : on assiste à une recherche de monopole impliquant des techniques spécifiques visant à imposer une définition *ad hoc* de la science mais aussi à verrouiller l'ensemble du champ scientifique (recrutements, formations, revues) et sa mise en débat public, comme cela a été montré ailleurs (Bédécarrats *et alii*, 2018 ; Jatteau, 2016). Ce verrouillage rend difficile toute forme de critique, puisque celle-ci s'apparente désormais à une prise de risque. Elle permet donc aux chercheurs de J-PAL de s'affranchir de toute controverse. La fabrique de CDDP (et du numéro spécial d'AEJ: AE), illustre cette position de domination, et y contribue. Elle leur permet aussi de faire fi du savoir accumulé dans le domaine des enquêtes, se traduisant par un manque patent de compétences qui explique une partie des erreurs observées. Cette lacune permet de faire fi des entorses multiples au protocole, qui pourtant entachent sérieusement la validité de l'étude. Elle leur permet de faire fi des critiques de l'AFD, quand bien même celles-ci ont été exposées publiquement et publiées. Elle leur permet enfin de continuer à prétendre que la méthode expérimentale est simple, pure et ce faisant supérieure à toutes les autres, quand bien même le protocole final révèle diverses erreurs de sondage et une complexité insoupçonnée.

Comme nous l'avons souligné nos travaux antérieurs (par exemple, Bédécarrats *et alii*, 2018), notre propos n'est pas de discréditer la méthode des RCT, mais de la reconnaître à sa juste valeur en contestant le piédestal sur lequel elle est aujourd'hui assise. Si les RCT restent probablement adaptées et légitimes pour certaines politiques circonscrites de manière précise, encore convient-il de les mener dans les règles de l'art. Il est à la fois nécessaire et possible d'employer d'autres méthodes sans rien concéder à la rigueur scientifique. Comme on l'a vu ici, ce pluralisme devrait s'imposer, notamment en complément des RCT afin de les contextualiser, tant en amont de la collecte qu'au moment de l'analyse. Il s'impose également pour toutes les questions, projets et politiques de développement qui ne rentrent pas dans le champ des RCT, et dont le microcrédit, caractérisé par des interventions pourtant relativement bien ciblées, est un exemple compte tenu de la faible participation et de la complexité de ses effets (Bernard *et alii*, 2012). Pour nombre de randomistes malheureusement, et J-PAL en particulier, « *les RCT ne sont pas seulement en tête de liste des méthodes, il n'y a rien d'autre au menu* » (Ravallion 2018, p. 10, notre traduction).

Combien de temps le piédestal peut-il tenir, et quand la bulle va-t-elle éclater (Picciotto, 2012) ? On pourrait rétorquer que le cas marocain est singulier et anecdotique. Il convient donc de multiplier les études de réplication et des approches plus fouillées de la fabrique des RCT comme celle-ci. Bien au-delà de notre propre analyse, d'autres positions critiques existent, dont certaines par des économistes de renom comme les Nobel d'économie James Heckman et Angus Deaton. Compte tenu des forces en présence, leur voix est précieuse. Plus globalement, c'est l'ensemble du champ scientifique qui mériterait d'être passé au crible. Les dérives observées ici ne sont-elles pas, *in fine*, l'expression d'un malaise grandissant au sein du champ académique, qui parvient de moins en moins à faire respecter les règles de base de l'éthique et de la déontologie ?

Liste des références

Akrich M., Callon M., Latour B. (dir.), (2006), *Sociologie de la traduction : textes fondateurs*, Presses des Mines, Paris.

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), pp. 151-82

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

Armendáriz B., Morduch J. (2010), *The Economics of Microfinance*, MIT Press (2ème édition), Cambridge, MA.

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), pp. 90–122.

Augsburg B., De Haas R., Fitzsimons E., Harmgart H., Meghir C. (2015), "The Impacts of Microcredit: Evidence from Bosnia and Herzegovina", American Economic Journal: Applied Economics, 7(1), pp. 183-203.

Bachelard G. (1965), *L'activité rationaliste de la physique contemporaine*, Presses Universitaires de France, Paris.

Banerjee A., Chandrasekhar A.G., Duflo E., Jackson M.O. (2013), "The Diffusion of Microfinance", *Science,* 341(6144), [DOI: 10.1126/science.1236498]

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

Banerjee A., Duflo E., Goldberg N., Karlan D., Osei R., Parienté W., Shapiro J., Thuysbaert B., Udry C. (2015b), "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries", *Science*, 348(6236), pp. 1–16.

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

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

Bédécarrats F., (2013), *La microfinance. Entre utilité sociale et rentabilité financière*, L'Harmattan, Paris.

Bédécarrats F., Guérin I, Morvant-Roux, S., Roubaud F. (2019), "Estimating microcredit impact with low take-up, high contamination and inconsistent data?", *International Journal for Re-Views in Empirical Economics* (à paraître).

Bédécarrats F., Guérin I, Roubaud F. (2018), "All That Glitters Is Not Gold. The Political Economy of Randomised Evaluations in Development", *Development and Change* [https://doi.org/10.1111/dech.12378]

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

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

Bouquet E., Wampfler B., Ralison E. et Roesch M., (2007), « Trajectoires de crédit et vulnérabilité des ménages ruraux : le cas des Cecam de Madagascar », *Autrepart*, vol. XLIV, pp. 157-172.

Bouhjar A. (2012), « De la langue du terroir à une langue au pouvoir : officialisation de la langue amazighe (berbère) au Maroc », *Lengas* [En ligne], 71 | 2012, mis en ligne le 07 avril 2014, consulté le 11 novembre 2018. URL : http://journals.openedition.org/lengas/343 ; DOI : 10.4000/lengas.343]

Bourdieu P. (1976), « Le champ scientifique », *Actes de la recherche en sciences sociales*, 2 (2-3), pp. 88-104;

Callon M. (2006), « Quatre modèles pour décrire la dynamique de la science », in Akrich M., Callon M., Latour B. (dir.), *Sociologie de la traduction : textes fondateurs*, Presses des Mines, Paris, pp. 201-251.

Cartwright N. (2007), *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, Cambridge, NJ.

Chemin M. (2008), "The benefits and costs of microfinance: Evidence from Bangladesh", *Journal of Development Studies,* 44, pp. 463–484.

Cheston S., Reed L. (1999), "Measuring Transformation: Assessing and Improving the Impact of Microcredit," *Journal of Microfinance / ESR Review*, 1 (1), pp. 21-43.

Clémens M.A. (2017), "The Meaning of Failed Replications: A Review and Proposal », *Journal of Economic Surveys*, 31(1), pp. 326–42.

Coleman B. (1999), "The impact of group lending in Northeast Thailand", *Journal of Development Economics* 60(1), pp. 105–141.

Colussi T. (2018), "Social Ties in Academia: A Friend Is a Treasure", *Review of Economics and Statistics*, 100(1), pp. 45-50.

Copestake J. Bhalotra S. Johnson S. (2001), "Assessing the impact of microcredit: A Zambian case study", *Journal of Development Studies*, 34(4), pp. 81-100.

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), pp. 123–150.

Crépon B., Devoto F., Duflo E., Parienté W. (2011), "Impact of microcredit in rural areas of Morocco: Evidence from a Randomized Experiment in Morocco," *MIT Working Paper*, 2011/3/31.

Cull R. J.; Morduch J. (2017), "*Microfinance and economic development"*, *Policy Research Working Paper*, No. WPS 8252, World Bank Group, Washington, D.C.

Deaton A. (1997), *The analysis of household surveys: a microeconometric approach to development policy*, The John Hopkins University Press, Baltimore et Londres.

Deaton A. (2010), "Instruments, Randomization and Learning about Development", *Journal of Economic Literature*, 48(2), pp. 424–55.

Deaton A., Cartwright N. (2018), "Understanding and Misunderstanding Randomized Controlled Trials", *Social Science & Medecine*, 210, pp. 2-21.

D'Espallier B., Labie M, Louis Ph. (2015), "Microcredit crises and unsustainable growth: A management perspective", *in* Guérin I., Labie M., Servet J.-M., *The Crises of Microcredit*, Zedbooks, Londres, pp. 23-38.

Desrosières A. (2013), *Gouverner par les nombres*, Presses de l'école des Mines, Paris.

Donovan K. P. (2018), “The rise of the randomistas: on the experimental turn in international aid”, *Economy and Society*, *47*(1), pp. 27–58. https://doi.org/10.1080/03085147.2018.1432153.

Duflo, E., P. Dupas and M. Kremer (2015), “School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools”, *Journal of Public Economics*, 123, pp. 92–110.

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, DOI: 10.1080/00220388.2011.646989

Duvendack M., Palmer-Jones R., Reed Robert W. (2017), “What Is Meant by ‘Replication’ and Why Does It Encounter Resistance in Economics?”, *American Economic Review*, 10 (5), pp. 46–51.

Eble A., Boone P., Elbourne D. (2014), “Risk and Evidence of Bias in Randomized Controlled Trials in Economics”, *Unpublished Working Paper*, Brown University, Providence, RI.

Fourcade M., Ollion E., Algan Y. (2015), “The Superiority of Economists”, *Journal of Economic Perspectives*, 29(1), pp. 89-114.

Garchitorena A., Murray M., Hedt-Gauthier B., Farmer P., Bonds M. RCTs in global health: contributions and unintended consequences, *in* Bédécarrats F. Guérin I. Roubaud f. (eds), *Randomized Control Trials in the Field of Development: The Gold Standard Revisited*, Oxford University Press, Londres (à paraître).

Grosh M., Glewwe P. éds. (2000), *Designing household survey questionnaires for developing countries : lessons from 15 years of the Living Standards Measurement Study : Volume One*, The World Bank, Washington D.C.

Guérin I., Morvant-Roux S., Roesch M., Moisseron J.-Y., Ould Ahmed P. (2011), « Analyse des déterminants de la demande de services financiers dans le Maroc rural », *Séries Analyse d’impact*, No.6, AFD, Paris.

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

Heckman J.J. (1991), “Randomization and Social Policy Evaluation”, *NBER Technical Working Paper* No. 107, National Bureau of Economic Research, Cambridge, MA.

Heckman J. J., Moktan S. (2018), “Publishing and Promotion in Economics: The Tyranny of the Top Five”, *NBER Working Paper No. 25093*, National Bureau of Economic Research, Cambridge, MA.

Hulme D., Mosley P. (1996), *Finance Against Poverty. Volumes 1 and 2*, Routledge, Londres.

Krauss A., (2018), “Why all randomised controlled trials produce biased results”, *Annals of Medicine*, 50(4), pp. 312−322.

Jamison J.C. (2017), “The Entry of Randomized Assignment into Social Sciences’. Policy Research Working Paper No. 8062. Washington, DC: World Bank.

Jatteau A. (2018), “Comment expliquer le succès de la méthode des expérimentations aléatoires ? Une sociographie du J-PAL », *SociologieS* [mis en ligne le 13 mars 2018, consulté le 19 octobre 2018].

Jatteau A. (2016), *Faire preuve par le chiffre ? Le cas des expérimentations aléatoires en économie*, thèse de doctorat, Université Paris-Saclay, Ecole Normale Supérieure, Paris.

J-PAL et IPA (2015), *Where credit is due*, Policy Bulletin, Cambridge MA. [https://www.povertyactionlab.org/sites/default/files/publications/where-credit-is-due.pdf (consulté le 21 novembre 2018].

Karlan D., Zinman C. (2011), “Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation”, *Science*, 332(6035), pp. 1278-1284.

Labrousse A. (2017), “Learning from Randomized Controlled Experiments. The Narrative of Scientificity, Practical Complications, Historical Experience”, *Books & Ideas*, Janvier : http://www.booksandideas.net/Learning-from-Randomized-Controlled-Experiments.html.

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*, n° 7, printemps, 32 p. : http://regulation.revues.org/7818

Latour B., Woolgar S. (1996), *La vie de laboratoire. La production des faits scientifiques*, La découverte (1ère édition 1979, trad. M. Biezunki), Paris.

Mix Market (2005), Mix 2005- Top 100 des IMF dans le Monde, [https://www.themix.org/sites/default/files/publications/MIX_2005_Global_100_MFI_League_Tables_FR_0.pdf, consulté le 12/12/2018.

Morvant-Roux S., Guérin I., Roesch M. Moisseron J.-Y. (2014), “Adding Value to Randomization with Qualitative Analysis: The Case of Microcredit in Rural Morocco”, *World Development*, 56, pp. 302–312.

Morduch J. (1998), Does microfinance really help the poor? New evidence from flagship programs in Bangladesh, New York University. [Retrieved from j.mp/bC3Tge]

Morduch J. (1999), “The Microfinance Promise”, *Journal of Economic Literature*, XXXVII (4), pp. 1569-1614.

Naudet J., Delarue J., Bernard T. (2012), “Évaluations d'impact : un outil de redevabilité ? Les leçons tirées de l'expérience de l'AFD », *Revue d'économie du développement*, 20(4), pp. 27-48.

Nubukpo (2012), « Esther Duflo, ou “l’économie expliquée aux pauvres d’esprit” », Blog *Alternatives Economiques* L’actualité vue par Kako Kubukpo. https://blogs.alternatives-economiques.fr/nubukpo/2012/02/20/esther-duflo-ou-%22l%E2%80%99economie-expliquee-aux-pauvres-d%E2%80%99esprit%22>.

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

Ogden T. (2017), *Experimental Conversations: Perspectives on Randomized Trials in Development Economics,* The MIT Press, Cambridge, MA.

Olivier de Sardan J.-P. (1995), *Anthropologie* et *développement. Essai en socio-anthropologie du changement social*, Karthala-APAD, Paris.

Pamiès-Sumner S. (2015), “*Development Impact Evaluation, State of Play and New Challenges”*, A Savoir (collection AFD), Paris.

Picciotto R. (2012), “Experimentalism and Development Evaluation: Will the Bubble Burst?”, *Evaluation* 18(2), pp. 213–29.

Pitt, Mark M., 2014. “Re-re-reply to the impact of microcredit on the poor in Bangladesh: revisiting the evidence", *Policy Research Working Paper Series 6801*, The World Bank.

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), pp. 958–996. [https://doi.org/10.1086/250037]

Pritchett L., Sandefur J. (2015), “Learning from Experiments when Context Matters, *American Economic Review: Papers & Proceedings*, 105(5), pp. 471–475.

Ravallion M. (2009), “Evaluation in the Practice of Development”, *World Bank Research Observer* 24(1), pp. 29–53.

Ravallion M. (2018), “Should the Randomistas (Continue to) Rule?” *Working Paper 492*, Washington: Centre for Global Development.

Reille X. (2009), Essor, crise et redressement du secteur de la microfinance au Maroc, CGAP-Brief, décembre 2009, 4 pages, retrived from : http://www.cgap.org/sites/default/files/researches/documents/CGAP-Brief-The-Rise-Fall-and-Recovery-of-the-Microfinance-Sector-in-Morocco-Jan-2010-French.pdf.

Rodrik D. (2008) ‘The New Development Economics: We Shall Experiment, but how Shall We Learn?’. Paper presented at the Brookings Development Conference, The Brookings Institution, Washington, DC (29–30 May).

Roodman D. (2011), *Due Diligence: An Impertinent Inquiry Into Microfinance*, Center for Global Development.

Roodman D., Morduch J. (2014). The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence, *Journal of Development Studies*, *50*(4), pp. 583–604. https://doi.org/10.1080/00220388.2013.858122

Rozas D. (2014), “Ending the Microfinance crisis in Morocco: acting early, acting right”, IFC, Washington, 48 p., retrived from https://www.ifc.org/wps/wcm/connect/5e1e5a0047850bdba0d4f5299ede9589/IFC+Morocco+MicroFinance+Crisis+report.English.pdf?MOD=AJPERES

Spears D., Ban R. Cumming O., (2018), “Trials and tribulations: the Rise and Fall of the Sanitation RCTs”, *in* Bédécarrats F. Guérin I. Roubaud f. (eds), *Randomized Control Trials in the Field of Development: The Gold Standard Revisited*, Oxford University Press, Londres (à paraître).

Sukhtankar S. (2017), “Replications in Development Economics”, *American Economic Review*, 107 (5), pp. 32–36.

Tarozzi A., Desai J., Johnson K. (2015) “The Impacts of Microcredit: Evidence from Ethiopia”, *American Economic Journal: Applied Economics*, 7(1), pp. 54-89.

Liste des documents non publiés mobilisés pour cet article

Interventions à des conférences ou séminaires

(I_1) : Bernard Tanguy, « Rural micro-finance in Morocco: Lessons from an on-going randomized study », Etude présenté à DIME Workshop, Dakar, févr. 2010 (23 slides).

(I_2): Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, « On measuring and bridging through impact evaluations: Lessons from AFD’s experience », présenté à NONIE Conference, Bonn, mars 2010 (12 slides).

(I_3) : Bernard Tanguy, « Impact evaluation of a micro-credit program in Morocco: A donor’s perspective », présenté à DIME Conference, Dakar, mai 2010 (24 slides).

(I_4) : Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, « Impact Evaluations and Microfinance Interventions: A donor’s perspective », présenté à CGAP Conference, Nairobi, mai 2010 (8 slides).

(I_5) : Bernard Tanguy, Delarue Jocelyne, Naudet Jean-David, « On measuring and bridging through impact evaluations: Lessons from AFD’s experience », présenté à EES Conference, Prague, oct. 2010 (12 slides).

(I_6) : Bernard Tanguy, « Mesurer et comprendre par les évaluations d'impact : Leçons d'expérience de l'AFD », Méthodologie présenté à Séminaire interne sur l'évaluation d'impact, Paris, déc. 2010 (12 slides).

(I_7) : Crépon Bruno, « Evaluer l'impact du micro-crédit en milieu rural », présenté à Présentation aux partenaires, Casablanca, mars 2007 (22 slides).

(I_8): Delarue Jocelyne, « Evidence and use: The impact evaluation of microfinance projects and their expected use », présenté à NONIE Meeting, Washington (DC), janv. 2008 (17 slides).

(I_9) : Delarue Jocelyne, « Impact evaluations from a bilateral donor's perspective », présenté à [Réunion à confirmer], Phnom Penh, juin 2009 (11 slides).

(I_10) : Delarue Jocelyne, « Les évaluations d'impact à l'AFD », présenté à CIRAD - GT Impact, Montpellier, mai 2010 (7 slides).

(I_11) : Naudet Jean-David, « We Shall Learn But Shall We Use? A Sponsor Perspective on Impact Evaluation »,], mars 2009 (10 slides).

(I_12) : Naudet Jean-David, « Evaluations d'impact et expérimentation : éléments de positionnement de l'AFD », présenté à Séminaire interne sur l'évaluation d'impact, Paris, déc. 2010 (8 slides).

(I_13) : Naudet Jean-David, « Tester ou évaluer les programmes de développement ? Quelques leçons d'expérience de l'AFD sur les évaluations d'impact », présenté à NONIE 2012, Paris, mars 2012 (11 slides).

(I_14) : Pamies Sumner Stéphanie, « Les évaluations d'impact dans le secteur de la microfinance à l'AFD : quelques retours d'expérience », présenté à Club microfinance Paris, Paris, mars 2012 (7 slides + compte rendu).

(I_15) : Pamies Sumner Stéphanie, « Impact evaluations: lessons from AFD's experience », présenté à SKY evaluation meeting, Phnom Penh, nov. 2011 (9 slides).

(I_16) : Pamies Sumner Stéphanie, « Les évaluations d'impact à l'AFD », présenté à Master 2 analyse de projets de développement durable, Rennes, 2012 (14 slides).

(I_17) : Pamies Sumner Stéphanie, « Impact evaluations at AFD: lessons learnt and perspectives », présenté à KfW Seminar, Franckfort, janv. 2013 (26 slides).

Documents internes : protocoles et amendements et courriers

(DI_1) : J-PAL (pas d'auteur), Evaluation de l'impact d'un programme de micro-crédit en milieu rural : Al Amana au Maroc, Projet de recherche soumis au Département de recherche de l'AFD, 21 janvier 2006, 39 pages.

(DI_2) : Réunion Copil 2007, Evaluation de l'impact du microcrédit en milieu rural, note préparée pour le comité de pilotage du 1ier août 2007, AFD, 16 pages.

(DI_3) : AFD, Compte rendu de mission sur les enquêtes de terrain de l'évaluation d'impact d'Al Amana, 29 avril 2008, 7 pages.

(DI_4) : Réunion Copil 2009, Comité de pilotage de l'Evaluation d'impact d'un programme de microcrédit en milieu rural au Maroc, minutes de la réunion du 29 janvier 2009, AFD, 4 pages.

(DI_5) : Lettre du directeur du département de la recherche de l'AFD à Mr. François Bourguignon, 19 mai 2008.

(DI_6) : Lettre de réponse du J-PAL au directeur du département de la recherche de l'AFD, 18 juillet 2008.