Anthropologie & développement

Anthropologie & développement

46-47 | 2017
Mythes sociotechniques et développement

Randomized Controlled Experiments to End Poverty?

A Sociotechnical Analysis

Nassima Abdelghafour

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/anthropodev/611
DOI: 10.4000/anthropodev.611
ISSN: 2553-1719

Publisher
Presses universitaires de Louvain

Printed version
Date of publication: 1 December 2017
Number of pages: 235-262
ISBN: 979-10-93476-05-6
ISSN: 2276-2019

Electronic reference
Nassima Abdelghafour, « Randomized Controlled Experiments to End Poverty? », *Anthropologie & développement* [Online], 46-47 | 2017, Online since 01 June 2018, connection on 10 December 2020. URL : http://journals.openedition.org/anthropodev/611 ; DOI : https://doi.org/10.4000/anthropodev.611

Randomized Controlled Experiments to End Poverty?

A Sociotechnical Analysis

Nassima Abdelghafour

*L'expérimentation aléatoire ou essai randomisé contrôlé (ERC) est une méthode d'évaluation inspirée des essais cliniques, transposée à l'économie du développement au début des années 2000. Rapidement devenue populaire, cette méthode est promue comme « l'étalon-or » de l'évaluation d'impact. Cet article examine un des premiers ERC, évaluant l'impact sur l'absentéisme scolaire d'un traitement vermifuge administré aux élèves dans une région rurale du Kenya. À travers l'étude de ce cas, il s'agit de mettre en évidence les processus par lesquels les ERC s'imposent comme une pratique d'évaluation incontournable. En insistant sur la production de résultats statistiquement non biaisés, les économistes défendant les ERC disqualifient les autres méthodes d'évaluation d'impact, et accentuent l'importance d'isoler l'impact causal d'une intervention de l'effet d'autres facteurs. Le type de preuves produites par les ERC engage ainsi un mode d'organisation des pratiques de lutte contre la pauvreté fondé sur la mise en compétition des interventions. Enfin, l'analyse d'une controverse scientifique venant remettre en question les résultats de l'expérimentation ouvre une discussion sur les politiques fondées sur les données probantes (*evidence-based policy*). Le poids de ces dernières n'est pas dû à une claire séparation entre science et politique, mais précisément à la manière dont science et politique sont entremêlées.*

*Randomized Controlled Trial (RCT) is an evaluation methodology imported from clinical trials to development economics in the early 2000s. It has been promoted as the "gold standard" of impact evaluation by its proponents. Focusing on the canonical case of the experiment testing the impact of deworming pupils on school attendance in a rural region of Kenya, this article inquires into the political success of RCT. By emphasizing the production of statistically unbiased results, RCT proponents disqualify alternative evaluation methods and stress the importance of attribution (*i.e.*,*

ensuring that the observed impacts are indeed attributable to the intervention and not to external factors). The type of evidence produced by RCT contributes to organizing competition between poverty-reduction interventions. Finally, the study of the "worm war", a scientific controversy challenging the results of the deworming experiment, leads to a discussion about evidence-based policy. I argue that evidence-based policy does not hold because of a clear separation between science and politics, but precisely because of the way they interplay.

Introduction[1]

> *"It's not the Middle Ages anymore, it's the 21st century. And in the 20th century, randomized controlled trials have revolutionized medicine by allowing us to distinguish between drugs that work and drugs that don't work. And you can do the same, randomized controlled trial for social policy. You can put social innovation through the same rigorous, scientific tests that we use for drugs. And in this way, you can take the guesswork out of policy-making by knowing what works, what doesn't work and why."*

These few sentences, taken from a TED talk[2] entitled "Social experiments to fight poverty", given by MIT economist Esther Duflo[3] in 2010, summarize a modernistic project for reforming anti-poverty interventions, through a systematic impact evaluation of social programs in the form of *in vivo* experiments in order to select the most effective programs for large scale implementation. Randomized controlled trials (RCTs) basically consist in comparing a group of *units* (*e.g*. individuals,

[1] Many thanks to Madeleine Akrich, Martin Denoun, Liliana Doganova, Vera Ehrenstein, Fiona Gedeon Achi, Anissa Pomiès, Vololona Rabeharisoa and the anonymous reviewers for helpful and insightful suggestions on the successive versions of this article.

[2] TED is a non-profit organizing, broadcasting and translating short and punchy talks on various topics, in order to globally spread ideas credited with a world-changing potential.

[3] Duflo is a leading figure of the RCT movement in development economics. She won many academic awards, among which the John Bates Clark medal, and is regularly praised in the media for her innovative approach to poverty.

schools, villages) receiving a *treatment*[4] with a group of units not receiving anything. The core assumption is that *random assignment* of units to treatment or *control* group ensures statistical similarity between the groups. In these conditions, any difference between them can be unambiguously attributed to the treatment, the impact of which can be estimated on an array of *outcomes* (*e.g*. health status, agricultural yields, income). The importing of the clinical trials methodology into the field of development economics was the initiative of a small team of economists that quickly grew into two connected, large and influential organizations, the Jameel-Poverty Action Lab (J-PAL), founded in 2003 and based at the MIT, and Innovations for Poverty Action (IPA), created in 2002 and based at Yale University. They have carried out hundreds of RCTs and actively publicized their methodology among academics, national and local governments, aid agencies, donors and the general public. RCT has been adopted by key actors of poverty reduction, both public agencies (*e.g*. DFID, USAID, UNICEF) and private donors (*e.g*. the Bill and Melinda Gates Foundation, the Hewlett Foundation).

RCT proponents defend a positivist vision of development based on faith in scientific and moral progress. Rather than taking part in the aid-effectiveness dispute opposing economists calling for increasing aid volumes (*e.g*. Jeffrey Sachs) and economists warning against the adverse effects of aid (*e.g*. William Easterly, Dambisa Moyo), RCT advocates propose breaking down this general theoretical problem into smaller practical problems. Their evidence-based approach, RCT-proponents argue, is free of ideology, dogmatic principles, political stances, and even free of theoretical assumptions about the nature of poverty. They pursue an ideal of objectivity. They claim that by generalizing the use of RCT, and through a trial-and-error process, a catalog of *best practices* can be put together in order to guide policy-making and drive funds towards *effective* and *cost-effective* projects. This methodic, iterative approach to poverty alleviation was initially presented as the antidote to the "guesswork"

[4] Italics signal expressions commonly used by economists doing RCT. Treatment, for example, is a term imported from clinical trials. It refers to the evaluated intervention, *i.e.* any social policy supposed to improve a given situation.

practiced by the World Bank, criticized by RCT proponents (Banerjee and He, 2003; Center for Global Development, 2006) for its lack of a consistent evaluation policy[5].

The displayed ambition for the generalization of RCT and the claim that RCT provides the best-quality evidence – its advocates refer to it as the *gold standard* – have not only provoked a controversy within the field of development economics but also attracted the scrutiny of social scientists from other disciplines. Authors argue that the hegemonic ambition of RCT is a problem, both practically and theoretically (Bédécarrats, Guérin and Roubaud, 2015). Contributions have questioned the validity of the reasoning underpinning RCT (Cartwright, 2007; Deaton, 2010), described the compromises made in practice between methodological rigor and practical implementation constraints (Quentin and Guérin, 2013), shown that RCT results are shaped by socio-political forces (Faulkner, 2014), highlighted the gap between the narrow scope to which RCT applies and the extensive use claimed by its advocates (Ravallion, 2012 ; Rodrik, 2008) and also specified the type of issues on which RCT can produce accurate knowledge (Bernard, Delarue and Naudet, 2012).

Let's focus on the political dimensions of RCTs. Their proponents have an agenda: they explicitly aim at transforming the international poverty-action scene by optimizing aid allocation (Banerjee and Duflo, 2011). However, they vigorously deny doing politics, and claim to consider the hard facts only, with no ideological or theoretical filter. This alleged neutrality has been challenged: RCT implicitly relies on theoretical corpuses (*e.g.* neoclassical micro-economics, experimentalism) that are not exempt from normative values (Durand and Nordmann, 2011; Picciotto, 2012). Even if we take seriously their effort to escape partisan debates on poverty, and their claim to rely solely on science to settle disagreements, it remains a very strong political gesture to depoliticize the issue of poverty. RCT has been characterized as an "evidence-based government" practice, where "the art of 'evidence-based government' is an art of emphasizing objectivity as a guarantee of realism and efficiency" and the concept of

[5] In reaction to these criticisms, the World Bank has since created an evaluation department which has conducted numerous RCTs.

government draws on Michel Foucault's work on governmentality (Bruno, 2015: 214). Also drawing on Foucault's work, authors have insisted on the paternalistic dimension of RCT (Labrousse, 2010), based on innovative forms of coercion (Bardet and Cussó, 2012).

In this contribution, I want to question the notion of "evidence-based policy", which is central to RCT. The phrase suggests that producing evidence is a first step, and political decision-making only comes afterwards, once evidence has been stabilized. I argue, on the contrary, that evidence and politics interplay throughout the production and circulation of evidence. Drawing on the sociology of translation[6], I analyze RCT as a sociotechnical device proposing, through its technical features, a vision of the world, as well as a form of social organization. Indeed, RCT is a complex device articulating techniques (*e.g.* data collection, logistics, computing, communication), theoretical corpuses (experimental sciences, social engineering, economics, inferential statistics), and material equipment (GPS, questionnaires, software) to produce a form of knowledge eventually materialized in academic papers, policy briefs, books and speeches. Technical artefacts in general propose a "script" (Akrich, 1991) that defines roles for users, distributes competences and organizes relations between people and their environment. Through the confrontation between this script and the environment where they are actually operated, technical devices produce a form of knowledge about the world while contributing to shape the world at the same time. If technical objects in general produce knowledge on their environment incidentally, RCT explicitly aims at doing so – which raises a series of specific issues around the political making of evidence and the political uses of such evidence.

I suggest an inquiry into the political success of RCT: what makes RCT attractive for major international development actors? I bring forward two types of explanation. First, I show that what makes RCT powerful is its capacity to exclude and make obsolete alternative practices of evaluation.

[6] Adapted to development studies, the sociology of translation provides an interesting vantage point to analyze the success or failure of development projects (Lavigne Delville, 2015 ; Le Meur, 2015 ; Mosse, 2005).

Then, I show that evidence-based policy does not hold because of a clear separation between science and politics, but precisely because evidence and politics are intertwined. The first section explores the "geography of competences"[7] (Akrich, 1991) organized by the experimental device. The second section describes the "gold standard" quality of RCT as the result of a specific framing of evaluation problems. The third section shows how RCT is used to compare various poverty-reduction interventions and follows the circulation of evidence along a network dedicated to *translating evidence into action*. The fourth section examines a scientific controversy, the "worm war", and questions the dynamics of evidence therein. The last section concludes.

I focus on the canonical experiment assessing the impact of deworming pupils on school attendance in Kenya. It is one of the first and most famous RCTs applied to development; it has led to massive deworming programs in several developing countries and is often used, by its advocates, as an example of how powerful RCT is. This case has been profusely documented. The analysis is based on a corpus of documents comprising academic papers, books and newspaper articles targeted at the general public, training material for students or for development professionals, blog posts, descriptions of experiments on the J-PAL and IPA's websites, policy briefs, and texts from the websites of nonprofits relying on RCT results to select the programs they implement or support.

What "geography of competences" is proposed by the experimental device?

Busia is a poor and densely populated rural district of Western Kenya, neighboring Lake Victoria. The Dutch nonprofit International Christian Support Fund Africa (ICS) has been operating in Busia since 1995, carrying out various interventions in local schools (*e.g.* distributing free uniforms, textbooks or flipcharts). In 1998, ICS launched the Primary School

[7] "Geography of competence" refers to the way technical decisions distribute competences across human actors and technical devices, and thus, to the way they contribute to organizing the environment.

Deworming Project (PSDP), covering 75 schools enrolling more than 30,000 pupils in total. The project took place in the southern part of the district, where intestinal worm infection rates are the highest (children get infected with worms when walking barefoot on contaminated soil)[8]. The deworming project was implemented in collaboration with the Kenyan Ministry of Health office in Busia (MHB), and evaluated by two development economists from the United-States: Ted Miguel and Michael Kremer (M&K), with funding from the World Bank and PSDP. Because of limited capacity, ICS could not reach all 75 schools at once. The necessity to gradually phase-in the program gave M&K the opportunity to implement a random assignment design for monitoring and evaluation. They estimated the impact of the deworming program on three arrays of outcomes: health, school attendance, and pupil performance. They published their findings in *Econometrica*, a prominent journal of economics (Miguel and Kremer, 2004). Had the experiment been limited to measuring the impact of the deworming treatment on health outcomes, it would have been quite similar to a clinical trial. But M&K tested a causal relationship between worm infection and school attendance. The question of interest is not to know how children's bodies will react to the drug, but rather, to understand how being dewormed will affect their social behavior.

Setting up an experiment to answer this question reveals a hypothetico-deductive understanding of the problem of school attendance. ICS and MHB, by launching the deworming program, formulated a hypothesis on Busia schoolchildren: they assume that deworming schoolchildren will improve their lives. ICS, MHB and M&K together reformulated this hypothesis by specifying the outcomes, for example: deworming schoolchildren will decrease school absenteeism. M&K were tasked to confirm or invalidate this hypothesis through quantitative analysis. ICS, MHB and M&K needed to negotiate a *modus operandi* allowing both ICS and MHB to deliver the program as they wish, and M&K to evaluate it. In their paper, M&K explain how schools are

[8] The intervention concerns two types of worms (geohelminths and schistosomiasis) that have different contamination patterns and require different medicine. However, the evaluation process is the same. I focus on the case of geohelminths, which are more widespread.

divided into three groups: schools are ranked alphabetically, then every third school is assigned to a different group. This is the result of a compromise:

> *"Private communication with Michael Kremer has confirmed that, in fact, the local partners would not permit the use of random numbers for assignment so that the assignment of schools to three groups was done in alphabetical order. […] Alphabetization may be a reasonable solution when randomization is impossible, but we are then in the world of quasi- or natural experiments, not randomized experiments" (Deaton, 2009).*

ICS, MHB and M&K cooperated closely to co-organize the experiment, but pupils and their families were not given an active part in this process. The experiment defines a "geography of competences" (Akrich, 1991) that denies the beneficiaries the reflexivity attributed to the other parties. Schoolchildren are given the passive role of the phenomenon to elucidate: they are expected to behave just as usual – they would not even need to know that there is an experiment going on to play their part in it. They are not associated to the reflection: they are not asked what prevents them from attending school. Their answers are considered less reliable than the result of an experiment:

> *"Speaking to [NGO workers and to the beneficiaries of the program] can uncover many stories of what is going on. […] But plausible explanations are not the same thing as answers." (J-PAL, n.d.)*

Moreover, asking them could be seen as influencing their behavior, and therefore biasing the experiment. Dialogue is seen as secondary to data collection, which can take the form of direct observations of the beneficiary (*e.g.*, fieldwork staff observes whether the child wears shoes or not) or of structured interviews. There is no place for unexpected discoveries in these surveys; they are used for quantitative analysis: the collected answers need to be easily and unambiguously coded and formatted into a dataset. The point is to describe a population, not to learn from people. Dialogue with the pupils and their families occurs on the margins of the experiment; it is not considered as the most relevant way to produce knowledge.

The initial hint that deworming might increase school attendance seems however to come from qualitative research. In their 2004 paper, M&K write: "nonexperimental studies suggest that worms do affect school participation" (Miguel and Kremer, 2004: 164), with proper reference in a footnote. The study they allude to investigates the way children in Western Kenya handle their health problems (Geissler *et al*., 2000). In this study, dialogue takes the form of interviews carried out by social scientists with children. In other cases, experimenters themselves have informal chats with poor people (Banerjee and Duflo, 2011). Qualitative research is put on the same level as anecdotal conversations[9]. Dialogue is neither formally part of the experiment nor recognized as a reliable source of knowledge. This raises the question of the conception of the treatment. Banerjee and Duflo (2009) observe (and welcome) the development of long-term partnerships between researchers and NGOs, which allows researchers to take a larger part in the framing of problems:

> *"In other words, the researcher was now being offered the option of defining the question to be answered, thus drawing upon his knowledge of what else was known and the received theory" (Banerjee and Duflo, 2009: 155).*

The economist's knowledge of the literature seems to prevail over the experience of local stakeholders (NGO workers and potential recipients).

But let's get back to the pupils and their families. Their part in the experiment is formalized in terms of *compliance* or *noncompliance*. Compliance means for pupils to act in conformity with their assignment to the treatment or control group. M&K estimate that 79% of the pupils assigned to treatment actually got treated in 1998 (and 59% in 1999). Children in the treatment group are supposed to take a deworming pill, but

[9] This raises the general problem of articulation between RCT and qualitative research (Jatteau, 2014; Labrousse, 2010; Quentin and Guérin, 2013). It also raises the delicate question of the relationships between economists and other social scientists. "The [economic] discipline's emphasis on mastering quantitative reasoning (widely interpreted as a sign of higher intellectual capabilities) certainly stands behind the often dismissive attitude of economists toward the other, less-formal social sciences" (Fourcade, Ollion and Algan, 2015: 90).

if they miss school on the deworming day, or if their parents do not give their consent, they are not compliant. This is the occasion to take notice that, in the experiment, deworming is school-based and randomized at the level of the school, not at the individual level. This design allows M&K to refine the economic analysis by measuring externalities (positive spillovers) of the treatment. It is also more convenient to implement and more acceptable for ICS and MHB: they might have been reluctant to randomize across individuals for ethical reasons. Nonetheless, it also has the effect of redistributing healthcare competences from the family sphere to public authorities (schools, local ministry of health office).

Of course, families are not completely excluded from the decision whether or not to deworm their child. But the (later-modified) consent rule in the first year of the experiment did not leave much room for noncompliance: parents opposed to deworming had to go and personally inform the school headmaster of their refusal, and were otherwise considered to be consenting to the treatment. Some noncompliance was induced by MHB nurses. The deworming protocol excluded girls over 13 from the treatment, even in treatment schools, because of a sanitary risk in case of pregnancy. This restriction gave M&K an occasion to measure within-school externalities – *i.e.* to see whether girls over 13 were positively affected by the fact that other children in their school were being dewormed. As worm infections are contagious, the idea is that even untreated children benefit from it, because they become less likely to be contaminated by other children. Some MHB nurses decided to deworm girls older than 13 anyway, estimating that the benefit outweighed the risk. These nurses, contrary to the other actors involved in the RCT, did not "subscribe" to this feature of the protocol (Akrich and Latour, 1992). Despite the efforts of the researchers and fieldworkers, the experimental protocol is but a proposition: if pupils, parents or nurses do not comply with it, all M&K can do is to estimate compliance rates and take them into account in their impact estimation strategy. The experimental device organizes an asymmetric geography of competences, but this initial "script" (Akrich, 1991) can be challenged when implemented on the field.

What makes RCT a "gold standard" according to its proponents?

What distinguishes RCT from other impact evaluation methods is that it allows to build a sophisticated counterfactual, *i.e*. a situation that simulates as credibly as possible what would have happened without the deworming program. "Counterfactual displays" can be defined as:

> *"how two future states of the world — one with the project and one without it — are played against each other and how the value of the project is derived from that interplay" (Ehrenstein and Muniesa, 2013: 162).*

These authors insist on the material dimension of counterfactuals:

> *"These do not rely solely on reasoning and imagination, but also require the production, circulation, and exhibition of documents and devices essential to valuation processes" (*ibid*.: 162).*

RCTs rely on a heavy material and logistic machinery, not only to implement the treatment, but also for data collection: producing statistical evidence requires data on a large population sample. In practice, teams of fieldworkers are brought to the field to survey people and enter data on computers. In the deworming experiment, 9,102 schoolchildren were interviewed, 2,328 provided a stool sample for parasitological diagnosis, and 778 got their blood tested for anemia.

How is the counterfactual built? M&K took advantage of the fact that ICS does not have the capacity of organizing deworming in all 75 schools in the same year. As already discussed, three groups of schools were constituted in a quasi-random manner. Group 1 schools received treatment in 1998, group 2 in 1999, and group 3 in 2001. There are two phases in this experiment: in 1998, group 1 is compared to groups 2 and 3, then in 1999 groups 1 and 2 are compared to group 3. Let's focus on the first wave, when group 1 schools (treatment) are compared to group 2 and group 3 schools (control). I focus on the most publicized result of the study: M&K found that deworming increases school attendance by 25%.

The impact of the treatment is estimated by comparing the variation in average school attendance in group 1 schools before and after the deworming campaign to the variation in average school attendance in group 2 and group 3 schools over the same period of time. The idea is that

children in groups 2 and 3 act just as children in group 1 would have acted without the deworming intervention. The key assumption is based on the statistical law of large numbers: beyond a certain sample size, random assignment ensures average similitude between all three groups. In other words, because of randomization, children in all three groups should have similar characteristics on average and are expected to react in a similar way to their environment. In this way, one can assume that the only difference between the three groups is the treatment and therefore that any difference in school attendance can be unambiguously attributed to deworming.

Indeed, many factors could influence school attendance and bias the estimated impact of the deworming intervention[10]. The idea behind RCT is that randomization and large numbers allow for these factors to be *controlled* for, absorbed by the control group, so as to isolate the pure impact of the intervention. Now of course, this is theory. After baseline survey data analysis, group 1 children were actually found to be worse-off than children in groups 2 and 3 on several health outcomes: randomization failed to produce three similar groups. This did not discourage M&K: they argued that the main risk associated to this bias is to underestimate the impact of deworming on school attendance. As they found a statistically significant impact of the treatment despite this initial bias, the unbalance between the groups is not treated as an issue in the paper.

Another practical difficulty is to measure school attendance:

> *"Since school attendance records are often poorly kept, school participation was measured during unannounced school visits by NGO field workers. Schools received an average of 3.8 school participation check visits per year in 1998 and 1999" (Miguel and Kremer, 2004: 189).*

[10] A frequently cited source of bias is the weather: rainfall can influence school participation in many ways, positively or negatively; it might increase the risk of malaria for example. If one group has better access to mosquito nets or chloroquine, children in this group will be less likely to be sick and to miss school. Then if all three groups have a similar access to mosquito nets and chloroquine, the effect of rainfall will be the same on average in all the groups and it will not show in the comparison.

Here, teachers are not entrusted with attendance recording competences, which are internalized in the experiment setting. But exact records for less than four days a year do not necessarily estimate attendance more accurately than poorly kept records for many days a year; in both cases, large measurement errors can be expected.

What RCT can theoretically do, and that other impact evaluation methods cannot do, is to estimate a statistically unbiased impact. That is its *gold standard* quality. RCT proponents constantly stress the importance of chasing down potential sources biases and other *threats to experimental integrity* when teaching or publicizing the methodology. Angus Deaton, a prominent RCT-skeptic (and Nobel-prize winner), argues in an interview that there is no compelling reason to prefer unbiasedness over other statistical qualities, in particular, precision:

> *"So a lexicographic preference for randomized control trials – the 'gold standard' argument – is sort of like saying we'll elevate unbiasedness over all other statistical considerations. Which you're taught in your first statistics course not to do. […] We often find a randomized control trial with only a handful of observations in each arm and with enormous standard errors. But that's preferred to a potentially biased study that uses 100 million observations. That just makes no sense"*[11] *(Ogden, 2017: 40).*

The communication effort made by RCT proponents to increase awareness about statistical biases contributes to make other evaluation methods obsolete. This distinctive characteristic of RCT – to be able to measure the pure impact of a program – altogether describes and performs (Mitchell, 2005) a world where attribution matters. Preference for unbiasedness is the result of a specific "problematization" process (Callon, 1986) that makes RCT the best evaluation solution.

[11] Fortunately, in many RCTs there are more than "a handful of observations in each arm" and the issue of precision is taken seriously. Nevertheless, Deaton's argument holds true; the very point of RCT is to produce unbiased results.

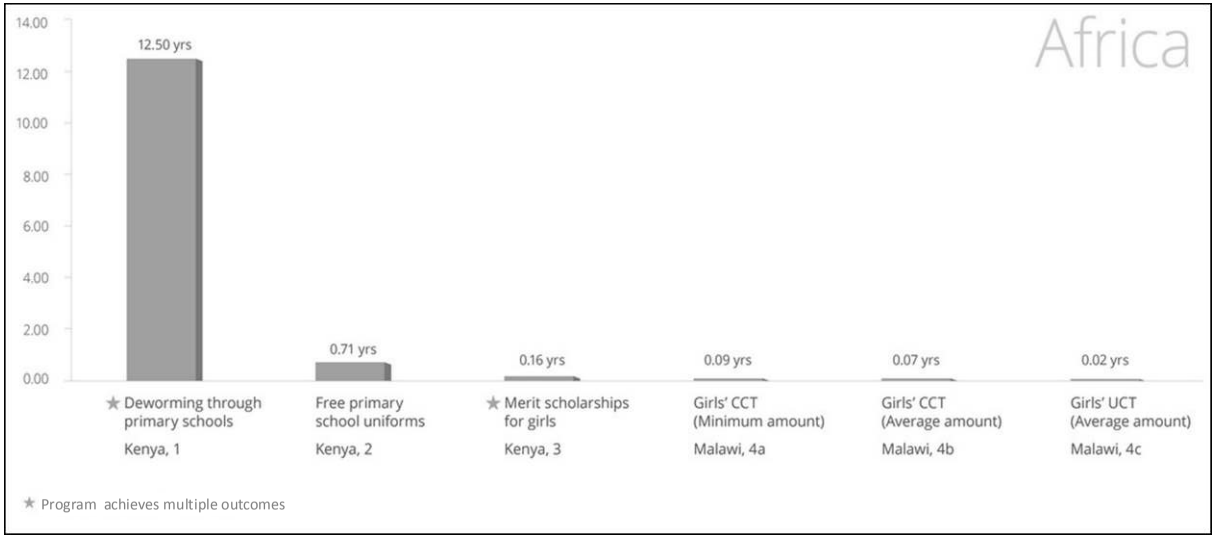

Figure 1. Cost-Effectiveness Analysis: Additional years of student participation per $100.

Source: https://www.povertyactionlab.org/policy-lessons/education/improving-student-participation

(CCT stands for conditional cash transfer and UCT for unconditional cash transfer)

How does RCT organize competition between poverty-reduction programs?

RCT stress the importance of knowing exactly which program is responsible for which outcome and forth, which organizations should get funding. Indeed, RCTs do not only ascertain whether a program works or not; they also provide a quantification of its impact. This allows to make several programs tackling the same issue commensurable, by comparing their cost-effectiveness ratios. Basically, it consists in dividing the impact of the program by its cost – of course, it is more complicated in practice (Dhaliwal *et al*., 2013). In Western Kenya, ICS, in collaboration with researchers affiliated to the J-PAL, tested several programs aimed at reducing school absenteeism. They tried to provide flipcharts (Glewwe *et al.*, 2000), to distribute free uniforms (Evans, Kremer and Ngatia, 2008), to offer a scholarship to high-performing girls (Kremer, Miguel and Thornton, 2004) and, of course, to deworm pupils. One of the methodological innovations that contributed to the success of the deworming study is that M&K took into account the *externalities* of the treatment. They estimated the impact of deworming on children who were not dewormed themselves, but who became less likely to get infected by their little neighbors. They showed that for each deworming pill distributed there is more than one child benefitting. Taking externalities into account decreases the cost of the program per person, which was already small, even without considering externalities. M&K conclude:

> *"Deworming was by far the most cost-effective method of improving school participation among a series of educational interventions implemented by ICS in this region of Kenya that were subject to randomized evaluations" (Miguel and Kremer, 2004: 205).*

The "policy lessons" pages of the J-PAL's website dedicated to the issue of school attendance compile the results of several RCTs carried out around the world. Results are broken-down per continent and presented in graphs. The graph for Africa is visually striking: deworming appears far more cost-effective than the other programs.

Cost-effectiveness analysis leads to reducing the various interventions carried out in different countries and at different times to only one dimension (here, years per student per $100 spent, Figure 1), making them

comparable. The argument of RCT proponents is a simple, basic economic argument: aid is a scarce resource that needs to be rationally and effectively allocated[12].

The evidence produced by RCT is a powerful mediation between academic researchers, development agencies, private foundations and NGOs. These figures are easily visualized on charts and graphs, and travel well. They can be described as a "metacode", or "pidgin trade language" shared by heterogeneous organizations (Rottenburg, 2009). This "metacode" facilitates the consolidation of a specific but wide-reaching network connecting organizations dedicated to *translating evidence into action*. Within this network, experimental results are turned into worthwhile causes. The deworming experiment, for instance, eventually gave rise to the "Deworm the World initiative". Massive school-based deworming programs have been organized, reaching over a hundred million children in Kenya, India and Ethiopia. IPA, who was initially in charge of the scale-up, finally created Evidence Action, a dedicated spin-off NGO, in 2013. Evidence Action benefitted from the support and endorsement of many other organizations. Deworming programs are for instance top-rated by GiveWell, a nonprofit organization that does "charity research" – the website uses the same visual codes as scientific journals and displays a very serious look. GiveWell could be described as a rating agency for the aid and philanthropy markets[13]. The information published on the website is supposed to help donors maximizing the impact of their philanthropic investment. Relying on systematic scientific literature reviews, GiveWell

[12] In this regard, the popularity of the deworming experiment was further strengthened by a follow-up article showing that dewormed children grow into more productive adults – interestingly, the outcomes highlighted by the authors reveal a belief in a very conventional path of development. "Ten years after the start of the program, the treatment group has better self-reported health, consume more meals, spend more time in entrepreneurship, and are more likely to grow cash crops. Kenyan women who participated in the program as girls have fewer miscarriages and reallocate labor time from agriculture to entrepreneurship. Men who participated as boys work 3.4 more hours each week, and are more likely to hold manufacturing jobs with higher wage earnings" (Baird *et al.*, 2015).

[13] GiveWell estimated its own impact at a total of 110.1 million dollars moved to its top-rated charities for the year 2015 (GiveWell, 2017a).

proposes a list of "top-charities" and establishes a ranking among them. It selects programs that are (supposedly) proven to have a large, positive and unambiguously attributable impact and to be highly cost-effective (GiveWell, 2017b). These criteria typically call for the type of results produced by RCT: quantitative evidence, cost-effectiveness ratios, statistically unbiased impacts. There is a strong tropism towards scientific literature exhibiting experimental results, and the selection of poverty-reduction programs is subsequently determined through this prism. In other words, GiveWell's ranking gives more information about which interventions are compatible with an evaluation by RCT, rather than about the interventions themselves.

Not only do RCTs discriminate between effective and non-effective programs; they also lead to the selection (and promotion) of so-called "best value for money" programs. If the use of RCT were to be generalized, there would be a risk of standardization of poverty-reduction policy through increased competition between programs. If, for each identified issue, there is a program labeled as the one maximizing the impact of the money spent, then why would a donor pick another program? The evaluated anti-poverty interventions are like black boxes that researchers are exempted from opening beforehand, because the experiment will conclude whether the intervention is effective or not. Once again, the process of (rigorous, scientific) evaluation seems to replace and disqualify (subjective, nonscientific, ideologically biased) discussion.

The "worm war" and the dynamics of evidence

The deworming experiment grew into a successful international program, and created a wave on which development economics is still surfing. Beyond the way this experiment tackled the particular issue of intestinal worm infection among school-age children, it also paved the way and provided a general roadmap for the production of further evidence-based poverty-reduction policy.

It did not go unchallenged though. A team of epidemiologists from the London School of Hygiene and Tropical Medicine used M&K's data and tried to replicate their findings. They went about it in two different ways.

First, they followed the same steps as M&K (Aiken *et al*., 2015). They took the computer program that was used in the original study and corrected errors in the code. This allowed them to identify many errors in the results, which M&K acknowledged – they had already found some themselves (Miguel and Kremer, 2014). But then, they also conducted their own analysis of M&K's data (Davey *et al*., 2015), with a different estimation strategy and based on different analytic choices – the use of statistics differs between economists and epidemiologists. They wrote their own code, based on a different interpretation of the experiment and on a different definition of the treatment. They ended up questioning the quality of the dataset, where a lot of variables were missing, challenging M&K's findings and, finally, contesting the size and robustness of the causal impact of deworming on school attendance.

These two publications started what was called the "worm war". Long and detailed articles proliferated on the development economics blogosphere[14], on social networks, and even in the generalist press. A heated dispute opposed those who claimed that deworming had been debunked by the epidemiologists, and those who accused Davey *et al.* of lacking elementary statistics skills, or of trying to create a buzz around their work. Many development economists ended up siding with M&K. Some of them (Chris Blattmann, Berk Ozler) even claimed to be even more convinced by the study than before its controversial replication. GiveWell adopted a median position. They acknowledged the errors found in the replication and the fact that these errors weakened the evidence provided by M&K. They even state some further reasons to be skeptical about M&K's findings. Nevertheless, they claim that deworming is still strongly supported by the long-term impact study by Baird *et al.*, which is, according to them, more convincing than M&K's study. They also argue that the very low cost of deworming balances the quality of the evidence:

> *"At the same time, because mass deworming is so cheap, there is a good case for donating to support deworming even when in substantial doubt about the evidence" (The GiveWell Blog, 2015).*

[14] David Evans provides a comprehensive list in a blog post: http://blogs.worldbank.org /impactevaluations/worm-wars-anthology

Deworming continued its career despite the controversy.

Who was right? The economists or the epidemiologists? Instead of trying to settle the dispute, let's focus on what the worm war teaches us about the political production of evidence. How can we explain the resilience of the global deworming project despite the controversy about the quality of the evidence supporting it? A first line of explanation draws on the sociology of translation (Callon, 1986). The strength of deworming lies in the network holding together various organizations which coordinate their actions around common principles of action (policy should be backed by "hard evidence") and common evaluation criteria (size and unbiasedness of impact, cost-effectiveness, potential for scale-up). Indeed, the notion of "hard evidence" efficiently translates the heterogeneous interests of these organizations into a common interest in supporting deworming programs. For the J-PAL, IPA, and development economists doing RCTs, deworming has become a flagship experiment, an example of what RCTs can do to guide poverty-reduction policy. For GiveWell, the type of evidence produced by RCT has rendered heterogeneous development interventions comparable, and has made ranking activities possible and relevant. For Evidence Action, the fact that deworming is supported by "hard" evidence is a way to leverage funding. For donors, the cost-effectiveness of deworming allows claiming a larger impact as well as a sound use of money.

A second possible explanation is related to the ability of RCT proponents to organize dissent among themselves. The re-analysis of M&K's dataset was indeed commissioned and funded by International Initiative for Impact Evaluation (3ie), a non-profit organization playing a strategic role in the promotion of evidence-based policy as a tool to reform development practices[15]. Contrary to many other actors who joined in the

[15] "3ie's Replication Program was established as a global public good to help improve the quality and reliability of impact evaluation evidence used for policymaking. The program is designed to highlight the benefits of internal replication of impact evaluations of development studies to the sector and to incentivize the conduct of replication studies of influential, innovative, and controversial impact evaluations of development interventions" (International Initiative for Impact Evaluation, 2016). 3ie affiliates include governmental agencies of developing countries (*e.g.* Planning Commission, Pakistan; Office of the Prime

worm war, 3ie did not aim at "debunking" (a word regularly used by the actors of the worm war) the analyses, but rather at emphasizing the fact that the original authors make their dataset and their code public. A call for re-analysis characterizes an impact evaluation as outstanding, transparent and replicable. In the "worm war" case, not only did M&K make a formal reply published on 3ie's website, but many other economists also got involved in the dispute, in a spirit of collective defense of their discipline (Allen and Parker, 2016).

Finally, the resilience of deworming projects may be due to a twofold line of argumentation that draws on the characteristic ambiguity of evidence-based policy. In M&K's 2004 paper, evidence clearly consists in a causal link articulating two different issues: worm infection and school attendance. Thus, the argument in favor of deworming is rooted in the correct articulation between these two issues: one should support deworming because it is a cheap and efficient way to increase school attendance and to boost human capital accumulation. But when the publication of a piece of counter-evidence challenged this causal link, the argument shifted onto the moral ground.

> *"We have made incredible progress over the past few years in getting more kids to have the chance to live worm-free lives. We cannot let weak scholarship and a flawed peer review process – let alone expensive treatment strategies – get in the way of this and hurt kids in the poorest countries around the world" (Evidence Action, 2015).*

In these two sentences, taken from a statement published by Evidence Action in the middle of the "worm war", the seriousness of the issue tackled (children's health) dramatizes the importance of producing good-quality evidence (as opposed to the so-called "weak scholarship"). But the issue of deworming is valuated *per se* and no longer because of its impact on other outcomes. It is regarded as morally good and desirable that

Minister, Uganda); public development agencies of developed countries (*e.g.* DFID, USAID); NGOs (*e.g.* Save the Children); development banks (*e.g.* the African Development Bank); research organizations (*e.g.* the J-Pal, the Institute for Development Studies) and private foundations (*e.g.* the Hewlett Foundation, MasterCard Foundation).

children can live without intestinal worms, and thus it is considered irresponsible to endanger deworming programs. The "incredible progress" that has already been accomplished supports the claim that deworming must continue: action is translated into evidence.

The initial trust in scientific evidence creates a strong attachment to the deworming program, which eventually contributes to give the deworming program some autonomy with respect to scientific considerations. The "worm war" raises the question of the dynamics of evidence. Initially, the legitimacy of deworming laid exclusively in the scientific credit of RCT. But then, this piece of evidence made its own way. Deworming developed, enrolled more and more organizations and materialized into a large network connecting nonprofit organizations, donors and national governments of several countries. By the time Davey *et al.* published their results challenging M&K's evidence, deworming had already gained momentum. The construction of large and complex sociotechnical networks transforming evidence into policy eventually makes poverty-reduction interventions less sensitive to counter-evidence, and goes against the trial-and-error spirit promoted by RCT advocates.

Evidence-based policy, politics of evidence

It seems that there are few limits to the expansion of RCT. The J-PAL alone has already organized 729 experiments in 67 different countries in Africa, Asia, Europe and the Americas. Expansion is not just geographic: RCTs evaluate more and more complex treatments. They do not only cover topics usually associated with poverty (*e.g.* health, education, agriculture, microfinance), but also issues such as governance, job market, corruption, political participation and crime. With a minimal theoretical toolbox composed of statistics and behavioral economics, RCT addresses a very large scope of issues. This gradual shift from the issue of poverty to other fields of applications of social engineering can be seen as a manifestation of the "imperialistic expansion of economics into aspects of social science that were traditionally outside the economic canon" (Fourcade, Ollion and Algan, 2015: 91). A systematic analysis of the programs evaluated through an RCT could bring valuable insights on the elusive politics of this device.

For now, let's build on what we have learnt by studying the experimental device.

RCT relies on a hypothetico-deductive understanding of poverty, and assigns poor people to a passive role in the experiment. The experimental approach, initially developed to investigate natural phenomena, problematizes poverty as an ahistorical and non-systemic issue that does not need to be related to previous events or states of the world, nor be understood as embedded in a global order. Hence, it circumvents north-south relationships and macroeconomic policy as potential explanations for poverty, and considerably reduces the subversive potential of anti-poverty action. Because comparison is core to RCT, potential solutions to poverty are framed as micro-level interventions targeting individuals (as opposed to nation-wide policies or to the provision of large infrastructure). This non-subversive, evidence-based approach to poverty enrolled a large constellation of actors coordinating their action around common principles and criteria. Indeed, the final users of RCT are not the poor participating in the field experiments, but the various actors who need what RCT produces, *i.e.* quantitative evidence. Despite controversies, RCT has imposed itself as the best way to evaluate the impact of a poverty-reduction program in major development institutions. Through a cumulative evaluation process, RCT modifies the ecology of anti-poverty practices and contributes to shape a world where poverty-reduction policy is standardized, rarefied, organized around a few established *best-practices*, at the expense of a diversity of interventions.

In conclusion, let's go back to our initial interrogation: how has RCT imposed itself as the "gold standard" of impact evaluation? Rather than providing firm answers, let's consider some elements for further discussion. David Mosse (2005) argues that "the order of development is primarily an interpretive order", meaning that development actors put more effort in trying to secure a particular interpretation of events than in trying to have actual control over the events themselves. Even though RCT was precisely promoted as an effort to shift the order of development onto the ground of events, facts and evidence rather than interpretations, the belief in evidence is itself an interpretation of the world. The global standardization of the economic profession and the "ideal of a 'monoeconomics,' tool-centered knowledge relatively insensitive to

historical and geographical variations" (Fourcade, 2006: 160) certainly makes it easier for economists to build a strong global interpretive community around the superiority of RCT. Another possible explanation for the success of RCT may be rooted in the efficacy of what Tania Murray Li (2007) calls "rendering technical":

> *"Contemporary development experts […] devise ever more restricted, technical interventions like giving children vitamins or deworming pills, and measure the outcome in terms of indicators like school attendance. They do not engage in debate over different possible futures, since the market can be counted on to direct human affairs efficiently and there is no alternative to it, or so we are told" (Li, 2015: 13).*

Thus, "rendering technical" goes together with "rendering non-political" – or more accurately, it makes the political dimension of development invisible. If one thinks of RCT as the sum of particular experiments, the operation of "rendering technical" provides each particular problematic situation with one indisputable best course of action. Now if one thinks of RCT at a more general level, as one sociotechnical device, it suggests something slightly different. One of the teachings of the "worm war" is that RCT reduced to its technical dimension – a standardized production process of quantitative evidence – is vulnerable to criticism. Indeed, quantitative evidence (namely, size and statistical significance of impact, cost-effectiveness) is produced after a complex, error-prone data collection process, and through analytical choices that can be challenged. RCT has imposed itself because it is promoted as a cutting-edge technical device and, in the same time, as a moral enterprise of helping the poor while making a rational use of aid money. The resilience of RCT draws on the mutual reinforcement of policy by evidence and of evidence by politics.

Bibliography

AIKEN A.M., DAVEY C., HARGREAVES J.R., HAYES R.J., 2015, "Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication", *International Journal of Epidemiology*, 44(5): 1572–1580.

AKRICH M., 1991, "L'analyse socio-technique," in VINCK D. (ed.), *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*, Bruxelles, De Boeck: 339–353.

AKRICH M., LATOUR B., 1992, "A Summary of a Convenient Vocabulary for the Semiotics of Human and Nonhuman Assemblies," in *Shaping technology, building society*, MIT press, Cambridge, Mass.

ALLEN T., PARKER M., 2016, "Deworming Delusions? Mass Drug Administration In East African Schools", *Journal of Biosocial Science*, 48(1): 116–147.

BAIRD S., HICKS J.H., KREMER M., MIGUEL E., 2015, "Worms at work: Long-run impacts of a child health investment", NBER working paper, 21428, National Bureau of Economic Research.

BANERJEE A.V., DUFLO E., 2009, "The Experimental Approach to Development Economics", *Annual Review of Economics*, 1(1): 151–178.

BANERJEE A.V., DUFLO E., 2011, *Poor economics: a radical rethinking of the way to fight global poverty*, 1st ed., New York, PublicAffairs: 303 p.

BANERJEE A.V., HE R., 2003, "The World Bank of the Future", *American Economic Review*, 93(2): 39–44.

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

BÉDÉCARRATS F., GUÉRIN I., ROUBAUD F., 2015, "The gold standard for randomised evaluations: from discussion of method to political economy", working paper, Paris, UMR DIAL.

BERNARD T., DELARUE J., NAUDET J.D., 2012, "Impact evaluations: a tool for accountability? Lessons from experience at Agence Française de Développement", *Journal of Development Effectiveness*, 4(2): 314–327.

BRUNO I., 2015, "Défaire l'arbitraire des faits. De l'art de gouverner (et de résister) par les 'données probantes'", *Revue Française de Socio-Économie*, Hors-série, 2: 213–227.

CALLON M., 1986, "Éléments pour une sociologie de la traduction. La domestication des coquilles Saint-Jacques et des marins-pêcheurs dans la baie de Saint-Brieuc", *L'année sociologique,* 36: 169–208.

CARTWRIGHT N., 2007, "Are RCTs the gold standard?", *BioSocieties*, 2(1): 11–20.

CENTER FOR GLOBAL DEVELOPMENT (ed.), 2006, *Rescuing the World Bank: a CGD working group report and selected essays*, Washington, DC, Center for Global Development, 201 p.

DAVEY C., AIKEN A.M., HAYES R.J., HARGREAVES J.R., 2015, "Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial", *International Journal of Epidemiology*, 44(5): 1581–1592.

DEATON A., 2010, "Instruments, randomization, and learning about development", *Journal of economic literature*, 48(2): 424–455.

DEATON A.S., 2009, "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development", working paper, 14690, Cambridge, Massachusetts, United States, National Bureau of Economic Research.

DHALIWAL I., DUFLO E., GLENNERSTER R., TULLOCH C., 2013, "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education", in *Education Policy in Developing Countries*, Chicago, University of Chicago Press.

DURAND C., NORDMANN C., 2011, "Misère de l'économie du développement", *La Revue des livres*, 1: 23–29.

EHRENSTEIN V., MUNIESA F., 2013, "The Conditional Sink: Counterfactual Display in the Valuation of a Carbon Offsetting Reforestation Project", *Valuation Studies*, 1(2): 161–188.

EVANS D., KREMER M., NGATIA M., 2008, "The impact of distributing school uniforms on children's education in Kenya", Washington, DC, United States, World Bank, mimeo.

EVIDENCE ACTION, 2015, “Worms Win, Kids Lose? Our Statement”, july 23, https://www.evidenceaction.org/blog-full/worms-win-kids-lose-our-statement

FAULKNER W.N., 2014, “A critical analysis of a randomized controlled trial evaluation in Mexico: Norm, mistake or exemplar?”, *Evaluation*, 20(2): 230–243.

FOURCADE M., 2006, “The construction of a global profession: The transnationalization of economics”, *American journal of sociology*, 112(1): 145–194.

FOURCADE M., OLLION E., ALGAN Y., 2015, “The Superiority of Economists”, *Journal of Economic Perspectives*, 29(1): 89–114.

GEISSLER P.W., NOKES K., PRINCE R.J., ODHIAMBO R.A., AAGAARD-HANSEN J., OUMA J.H., 2000, “Children and medicines: self-treatment of common illnesses among Luo schoolchildren in western Kenya”, *Social Science & Medicine*, 50(12): 1771–1783.

GIVEWELL, 2017, “GiveWell Metrics Report – 2015 Annual Review”, http://www.givewell.org/about/impact

GIVEWELL, 2017, “Our Criteria for Top Charities”, http://www.givewell.org/how-we-work/criteria

GLEWWE P., KREMER M., MOULIN S., ZITZEWITZ E., 2000, “Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya”, working paper, 8018, Cambridge, Massachusetts, United States, National Bureau of Economic Research.

INTERNATIONAL INITIATIVE FOR IMPACT EVALUATION, 2016, “Impact Evaluation Replication Programme”, http://www.3ieimpact.org/en/evaluation/impact-evaluation-replication-programme

JATTEAU A., 2014, “Expérimenter le développement ?”, *Genèses*, 93(4): 8–28.

J-PAL, n.d., “Introduction to Evaluations”, https://www.povertyactionlab.org/research-resources/introduction-evaluations

KREMER M., MIGUEL E., THORNTON R., 2004, “Incentives to Learn”, working paper, 10971, Cambridge, Massachusetts, United States, National Bureau of Economic Research.

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* [en ligne], 7.

LAVIGNE DELVILLE P., 2015, "30 mai 2015 – Comprendre le 'succès' et 'l'échec', lire les processus : l'apport de la sociologie de la traduction", https:// anthropo-impliquee.org

LE MEUR P.Y., 2015, "Un barrage contre le Pacifique: polders et développement au Cambodge", *Anthropologie & développement*, 42-43: 27–58.

LI T.M., 2007, *The will to improve: governmentality, development, and the practice of politics*, Durham, Duke University Press, 374 p.

LI T.M., 2015, "Governing rural Indonesia: convergence on the project system", *Critical Policy Studies*, 10(1): 79–94.

MIGUEL E., KREMER M., 2004, "Worms: identifying impacts on education and health in the presence of treatment externalities", *Econometrica*, 72(1): 159–217.

MIGUEL E., KREMER M., 2014, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, Guide to Replication of Miguel and Kremer (2004)", UC Berkeley, Center for Effective Global Action, 48 p.

MITCHELL T., 2005, "The work of economics: how a discipline makes its world", *European Journal of Sociology/Archives Européennes de Sociologie*, 46(2): 297–320.

MOSSE D., 2005, *Cultivating Development: An Ethnography of Aid Policy and Practice*, London; Ann Arbor, MI, Pluto Press, 315 p.

OGDEN T.N. (ed.), 2017, *Experimental conversations: perspectives on randomized trials in economic development*, Cambridge, MA, MIT Press.

PICCIOTTO R., 2012, "Experimentalism and development evaluation: Will the bubble burst?", *Evaluation*, 18(2): 213–229.

QUENTIN A., GUÉRIN I., 2013, "La randomisation à l'épreuve du terrain", *Revue Tiers Monde*, 213: 179–200.

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

RODRIK D., 2008, “The New Development Economics: We Shall Experiment, but How Shall We Learn?”, HKS Working Paper, No. RWP08-055, Cambridge, Massachusetts, United States.

ROTTENBURG R., 2009, *Far-fetched facts: a parable of development aid*, Cambridge, Mass, The MIT Press (Inside technology), 235 p.

THE GIVEWELL BLOG, 2015, “New deworming reanalyses and Cochrane review”, July 24, http://blog.givewell.org/2015/07/24/new-deworming-reanalyses-and-cochrane-review

Nassima Abdelghafour is PhD Candidate in Sociology

Centre de sociologie de l'innovation (Mines Paristech)

E-mail : nassima.abdelghafour@mines-paristech.fr