Papiers de Recherche | Research Papers

All That Glitters Is Not Gold:

The Political Economy of Randomised Evaluations in Development

Florent BÉDÉCARRATS
Isabelle GUÉRIN
François ROUBAUD

May 2017

Papiers de Recherche de l'AFD

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ils s'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l'AFD n'en exclut aucune autre.

L'Agence Française de Développement (AFD), institution financière publique qui met en œuvre la politique définie par le gouvernement français, agit pour combattre la pauvreté et favoriser le développement durable. Présente sur quatre continents à travers un réseau de 72 bureaux, l'AFD finance et accompagne des projets qui améliorent les conditions de vie des populations, soutiennent la croissance économique et protègent la planète. En 2014, l'AFD a consacré 8,1 milliards d'euros au financement de projets dans les pays en développement et en faveur des Outre-mer.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s).

Les *Papiers de Recherche* sont téléchargeables sur : http://librairie.afd.fr/

AFD Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

Agence Française de Développement (AFD), a public financial institution that implements the policy defined by the French Government, works to combat poverty and promote sustainable development. AFD operates on four continents via a network of 72 offices and finances and supports projects that improve living conditions for populations, boost economic growth and protect the planet. In 2014, AFD earmarked EUR 8.1bn to finance projects in developing countries and for overseas France.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

AFD Research Papers can be downloaded from: http://librairie.afd.fr/en/

All That Glitters Is Not Gold. The Political Economy of Randomised Evaluations in Development

Florent Bédécarrats, AFD Evaluation Unit.

Isabelle Guérin, IRD-CESSMA.

François Roubaud, IRD-DIAL.

Abstract

Randomised Control Trials (RCTs) have a narrow scope, restricted to basic intervention schemes. Experimental designs also display specific biases and political uses when implemented in the real world. Despite these limitations, the method has been advertised as the gold standard to evaluate development policies. This paper takes a political economy angle to explore this paradox. It argues that the success of RCTs is driven mainly by a new scientific business model based on a mix of simplicity and mathematical rigour, media and donor appeal, and academic and financial returns. This in turn meets current interests and preferences in the academic world and the donor community.

Introduction

This last decade has seen the emergence of a new field of research in development economics: randomised control trials (hereinafter referred to as RCTs). Although the principle of RCTs is not scientifically new, their large-scale use in developing countries is unprecedented. These methods borrowing from medical science were first put into use for public policy evaluation in developed countries back in the 1960s (mainly in the United States in areas such as criminology, insurance, employment, taxation and education; Oakley, 2000; Pritchett *et al.*, 2013). They have since been tailored to poor countries' issues and circumstances. RCTs have been a resounding success, as seen from their proliferation, and their promoters have quickly won international accolade. Leading economic journals have welcomed RCTs with open arms in a surge of published articles. Their reputation now transcends the disciplinary field of economics and extends as far as the prestigious *Science* magazine, which has opened its pages to RCTs (Banerjee *et al.*, 2015b). Rare are the academic courses professing to "teach excellence" today that do not include a specialised module in this field, as found in the leading American universities (Harvard, MIT, Yale, etc.), the London School of Economics and the Paris, Toulouse and Marseille School of Economics. Rare also are the international conferences that do not hold crowd-drawing sessions on RCTs. And rare are the aid agencies that have not created a special RCT department and have not launched or funded their own RCTs.

RCTs represent an indisputable advance in development economics methodology and knowledge. Yet despite their limited scope of application (evaluation of specific, local and often small-scale projects), RCTs are now held up as the evaluation gold standard against which all other approaches are to be gauged. Presented by their disciples as a true Copernican revolution in development economics, they are the only approach to be proclaimed "rigorous" and even "scientific". Some media celebrity RCT advocates, with Esther Duflo in the forefront, are looking to take RCTs well beyond their methodological scope in a move to establish a full list of good and bad development policies. The grounds put forward for this upscaling ambition are an ever-growing number of impact studies from which scalable lessons can be drawn. Clearly though, there are a certain number of drawbacks to the proclaimed supremacy of RCTs in quantitative evaluation, which will be discussed here: disqualification and crowding out of alternative methods, ever-growing use of allocated resources, and rent positions. There is hence a real gulf between their narrow scope and the supremacy claimed by the highest-profile promoters of RCTs. Such is the paradox we propose to explore in this paper.

In the first section, we briefly present the founding principles of RCTs with their theoretical advantages, especially compared with other existing evaluation methods, and their meteoric rise. The second section discusses and defines the real scope of these methods. It identifies their limitations, especially when used on the ground outside of their ideal laboratory conditions, and studies their political uses (and ends) in a move to establish the extent of their validity in the more general development arena. The third and last section looks at the method as a whole from a broader political economy angle. Political economy is defined here as the interplay between political forces (which may be institutions, organised groups or individuals) and economic

activities (in this case RCTs, which have become a real industry as we will show) and how these two aspects mutually influence one another. From this point of view, we seek to understand how the different players interact, the power games and balances, and who gains from them. This last section presents an explanation as to why these methods enjoy political and academic credibility far beyond their real scientific scope. The conclusion presents our own view of impact evaluation methods and some avenues of research to take forward this paper.

I.   The rise of a methodology

Randomised control trials are designed to compare the outcome of a project (programme or policy) with what would have happened without the intervention in order to measure its net impact, i.e. minus all the changes occurring elsewhere. The challenge is to build the baseline scenario (the project-free *counterfactual*) which, by definition, is never observed. The solution proposed by randomised control trials is to draw two samples at random from a population likely to benefit from the intervention. The project is allocated to just one of the groups, but surveys are conducted on both groups before and after the project. Statistical properties taken from survey sampling theory guarantee that, on average, the differences observed between beneficiaries and non-beneficiaries can be attributed to the project. As with all probabilistic methods, results are reported with a margin of error (confidence interval), which depends on the sampling characteristics (size, method, attrition, etc.).

Randomised control trials hence seek to formally establish a causal link between an intervention and a certain number of outcome variables. Scientifically, and theoretically, they could legitimately be said to be the most convincing option available to identify the existence and quantify the magnitude of the observed impact. In quantitative evaluations, they are arguably more robust than other methods: when a control group is not set up before the fact, the *before-after* and *with-without* approaches cannot control for changes in context; quasi-experimental matching methods – which match beneficiaries and non-beneficiaries based on shared observable characteristics – partially lift this constraint. However, without ex-ante random selection, they omit the unobservable characteristics that might have influenced selection and would therefore differentiate the treatment group from the control group (risk aversion, "entrepreneurship", inclusion in social networks, etc.). Again in quantitative evaluations, RCTs in principle meet the methodological challenge of demonstrating the direction of causality without relying on complex econometric and still-refutable assumptions. Last and more classically, they differ from qualitative methods (case studies, monographs, interviews and participant observation) in their quantitative measurement of the impact, which is beyond the reach (and purpose) of qualitative methods. RCTs have become such a must to estimate causal relations that, although initially at odds with econometric techniques, they have become their "benchmark" as evidenced by the title of the introductory chapter ("*The Experiment Ideal*") to a highly popular econometrics manual (Angrist & Pischke, 2009).

In addition to these generic (and theoretical) plus points, there are other reasons to commend the replication and proliferation of RCTs in development. We note the three main reasons. Firstly,

RCTs have put their finger on a blind spot in both national policies and policies driven by official development assistance (ODA) in developing countries, which is their glaring lack of quantitative evaluation in the past. Massive sums have been spent without any clear idea of policy effectiveness, leaving these policies wide open to severe criticism for ideological, more than scientific, reasons (Easterly, 2007; Moyo, 2009). Acceptance of the principle of evaluations and their proliferation can but contribute to democratic accountability in the South and the North (Cling *et al.*, 2003). Secondly, RCTs have ramped up first-hand survey data collection by development economists. Researchers, long restricted to modelling macroeconomic aggregates from huge international databases of dubious quality, especially in Africa (Jerven, 2013 & 2015; Devaradjan, 2013), can now take advantage of the credibility accorded RCTs by mainstream economics to plug into the grassroots level and stakeholders. Thirdly, economic research used to marginalise developing countries because they lacked quality data, especially longitudinal data. The widespread use of RCTs brings economic research on these countries up to world class level. It even stands as a methodological advance initiated in the South and transferred to the North.

In the mid-2000s, these advantages drove a staggering boom in RCTs in developing countries. The Abdul Latif Jameel Poverty Action Lab (J-PAL) was one of the most influential promoters of RCTs, spearheading a vast advocacy campaign for them. J-PAL was founded in 2003 by Massachusetts Institute of Technology researchers Abhijit Banerjee and Esther Duflo along with Harvard researcher Sendhil Mullainathan. The Lab works exclusively on RCTs and is a recognised quality label in the field. It organises training courses and exchanges of practices with a network of 146 affiliated professors[1] and a host of researchers (Figure 1). It helps find funding for randomised studies and promotes the dissemination of results to scientific circles and policymakers. The laboratory is closely associated with IPA (Innovations for Poverty Action), an NGO working to scale up the evaluations and programmes. In January 2017, thirteen years after its establishment, J-PAL was posting a total of no less than 811 evaluations (ongoing or completed) in 74 countries, with steady growth over the years. Africa is its leading ground (with 240 trials), way ahead of South Asia (165, mainly in India) and Latin America (131). Top focal points are finance (241) followed by the social sectors (196 for education and 166 for health) and governance, attracting exponential attention (185).[2] Esther Duflo plays a key role in the Lab, having worked on no less than 49 trials (13 of which are ongoing), but even she is largely outdone by Dean Karlan who has 100 trials (42 ongoing) to his name! Yet this very display of performance begs the question as to whether these researchers are really plugged into the grassroots level (we will come back to this).

Although the World Bank appears to have more of a mixed bag of impact evaluation methodologies (albeit always quantitative), RCTs represent nearly two-thirds of the methods used, accounting for 64% of the 368 evaluations undertaken (as at 2010). As RCTs shift into an increasingly dominant position, they are crowding out the other approaches, as shown by Figure 1. From 2000 to 2004, barely 20% of all evaluations were RCTs. The subsequent six-year

---

[1] Information collected from J-PAL's website on 24 January 2017.
[2] In the latest version of the evaluation database we looked up on the J-PAL website, one evaluation can cover more than one sector.

period saw a complete reversal of these proportions (76%). The number of RCTs is steadily climbing as evaluations based on other methods stagnate, if not backslide. Unfortunately, these consolidated figures for the World Bank have not been updated since 2010, but the situation is unlikely to have changed significantly. The establishment of DIME (Development Impact Evaluation Initiative; see Section III) in 2005 has been followed by the launch of other specialised RCT funds, such as the Strategic Impact Evaluation Fund (SIEF) in 2007, the Global Agriculture and Food Security Program (GAFSP, with up to 30% of its projects evaluated by trials) in 2009 and Impact Evaluation to Development Impact (I2I) in 2014. In 2016, Esther Duflo stated that approximately 400 ongoing projects financed by the World Bank were being evaluated by RCTs (Duflo, 2016).

Figure 1: Growth in J-PAL and World Bank impact evaluations
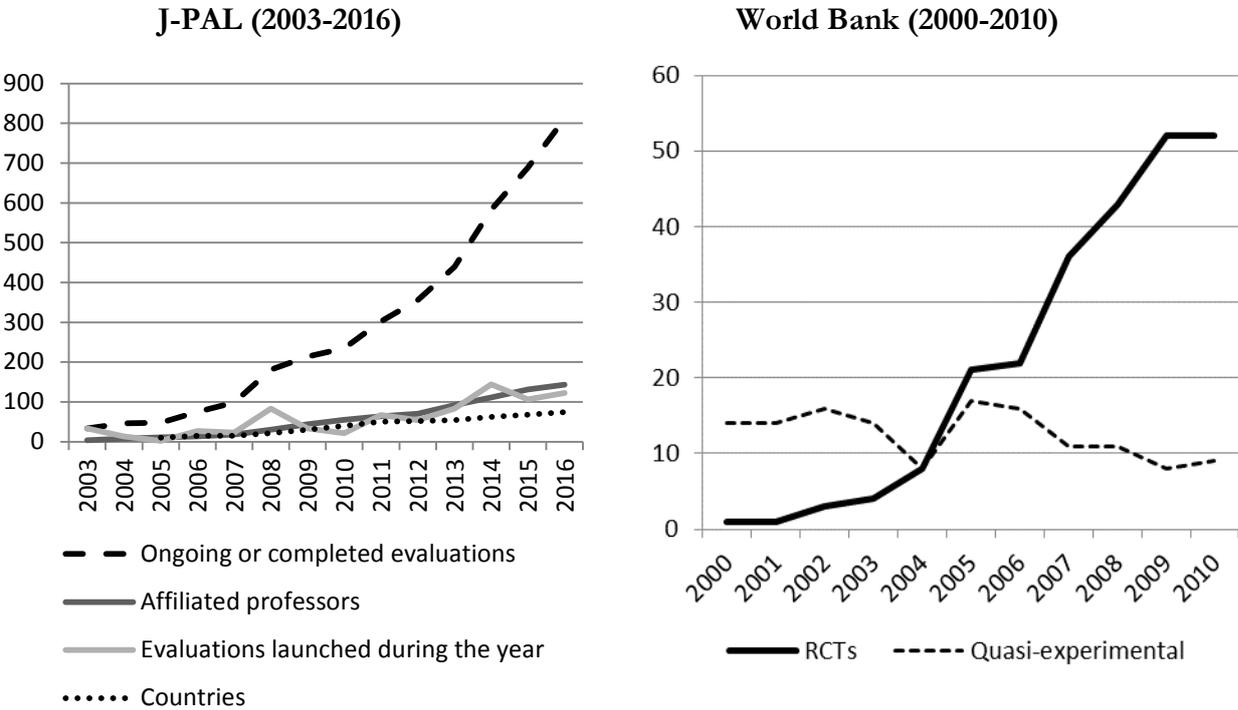
Source : Jatteau (2016) for 2003-2015, and J-PAL website for 2016 (updated 6 January 2017); IEG, 2012.

So impact evaluations have been a dazzling success story in recent years, to the point of becoming quite the industry. And RCTs have taken the lion's share: the August 2016 updated 3ie [3] Impact Evaluation Repository contains over 4,260 completed or ongoing development impact evaluations, 2,645 of which are RCTs (Miranda *et al.*, 2016; Figure 2). The question as to whether the golden age of the "gold standard" is over remains open (we will come back to this in the conclusion). Figure 2 presents a flattening RCT growth curve even though volumes remain high, while other indicators show no such stagnation (see Figure 1, for example). This raises the

---

[3] 3ie (International Initiative for Impact Evaluation) is an international NGO founded in 2008 and supported by leading foundations and donors. It promotes evidence-informed development policies and programmes in developing countries. As at September 2016, it had granted 108.8 million dollars in subsidies for studies, mainly to fund 216 impact evaluations (source: http://www.3ieimpact.org/en/about/performance-metrics/, updated in September 2016 and consulted on 2 February 2017).

question as to whether this apparent stagnation might not actually be the effect of an as yet incomplete inventory on the most recent years.

The RCT industry that has taken off in the last ten years remains a booming sector today, largely dominating the field of development policy impact evaluation. RCTs have spread to the four corners of the globe, rallying considerable resources. Data are not available to quantify the financial flows concerned, but it is sure they run to hundreds of millions of dollars. For example, although J-PAL does not publish financial reports, IPA's annual revenue rose from 252,000 dollars in 2003 to over 39 million dollars in 2015.[4] RCTs have generated a host of best practice manuals and dozens of academic papers, especially in leading journals: in 2015, RCTs accounted for 31% of the development economics articles published in the top five journals[5] and over 40% in the next tier of general interest journals[6] (McKenzie, 2016). They also guide careers: 65% of BREAD affiliated economists who have received their PhD since 2006 have worked on at least one RCT (Duflo, 2016). Yet why is this striking phenomenon (called a pro-impact evaluation movement) all the rage?

**Figure 2: Growth in the number of development impact evaluations
by type of method (1995-2015)**
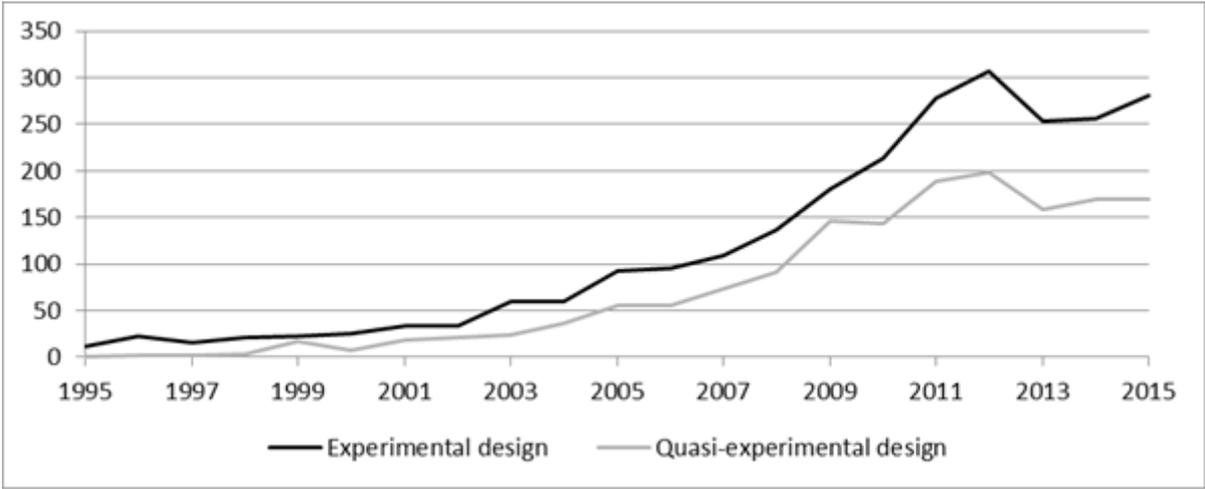
Source : 3ie Impact Evaluation Repository, updated in August 2016.

Is this systematic use of RCTs scientifically sound and politically expedient? Two questions pose problems: first, the proclaimed intrinsic superiority of RCTs over any other method (Duflo *et al.*, 2007), and second, the idea that the ubiquitous build-up of RCTs will, by sheer force of numbers, answer all development questions, about "what works and what does not work", based on indisputable foundations (Duflo & Kremer, 2005); claims that the movement's promoters have not since dropped. Karlan's testimony (2015) before the U.S. Congress is a perfect illustration of this.

---

[4] Biannual financial reports available at www.poverty-action.org/about/annual-reports-finances, consulted on 24 January 2017.
[5] IDEAS-RePEc ranks the top five as *The Quarterly Journal of Economics*, *Journal of Political Economy*, *American Economic Review*, *Econometrica* and *Journal of Economic Literature* (list as at 2 February 2017).
[6] *American Economic Journal: Applied Economics*, *Economic Journal* and *Review of Economics and Statistics*.

II.   Critiques of the method: from theory to implementation

Advocates of the use of RCTs in development economics imported the method from the medical world without due consideration of the critical discussions, conditions for their use and questions already raised about them in the public health sphere (Labrousse, 2010; Eble *et al.*, 2014). They also passed over the controversies that had marked decades of development economics debates (Picciotto, 2012). We summarise here the main critiques of RCTs, in terms of both their internal and external validity, before looking into the implications of their use on the ground. We take concrete examples to show that the political economy of RCTs has first-order consequences in practice, which call into question the method's asserted theoretical properties.

2.1.   Discussion of the (internal and external) validity of RCTs

The internal validity of RCTs (i.e. the reliability of the results obtained) is supposed to be the method's main strong point. Yet RCTs are far from free of limitations. Although their limitations are essentially empirical, in the way in which RCTs are conducted on the ground, some are more theoretical. The main criticism in this regard concerns the tension between bias (to be minimised) and precision (to be maximised), where RCTs fail to make an optimal trade-off in keeping with the basic principles of decision theory (Deaton & Cartwright, 2016).[7] Randomised analyses focus on average results for the entire population considered. Although RCTs in principle provide unbiased estimators of the average effect, they do not guarantee the minimal variance of these estimators any more than they can calculate the median effect or effect by quantile. This is particularly problematic when treatment effects are heterogeneous. Yet in medicine, and even more so in development, patients or users are liable to react in a wide range of different ways to a treatment. Where there is no information on variance, there is no robust method available to quantify the RCTs' margins of error and the tests used are actually more often than not inadequate. Precision could be improved by stratification methods, for example, but this implies a minimum of prior knowledge of the context, which many randomistas refuse to use. In a way, the assertion of the superiority of randomisation trips over its own feet: improved estimates call for the ability to make educated assumptions, whereas the RCT movement has established this method's superiority precisely on its absence of priors, using nothing more than the law of large numbers. Last but not least, RCTs do nothing to solve the classic problems of statistical inference (especially in the presence of asymmetric distribution, outliers, etc.). The conventional practice of departing from the simple comparison of treatment and control group means by estimating econometric regressions is not a solution; far from it, as shown by Young (2015). After applying a suitable test to 2,003 regressions in 52 RCT papers published in leading American journals, he finds that 30% to 40% of the coefficients found to be significant are actually not. All these problems combined lead Deaton and Cartwright (2016) to firmly conclude that, "It is almost never the case that an RCT can be judged superior to a well-conducted observational study simply by virtue of being an RCT."

---

[7] See also, among others, Heckman (1991), Deaton (2010), Ravallion (2009), Barrett & Carter (2010), and Rodrik (2008).

Empirically, RCTs are no exception to the habitual "tweaking" of research protocols, especially in social science. Many implementation constraints undermine the very premises of random sampling and hence the foundations of the purported scientific superiority of RCTs. Of note here is the absence of any real random sampling for ethical or practical reasons (so as not to disturb a project's implementation) (Scriven, 2008; Labrousse, 2010; Deaton, 2010); treatment variables often too poorly specified to be able to serve as structural estimators (Heckman & Vytlacil, 2005); the absence of blinding, which introduces different incentives for participation or non-participation (Heckman *et al.*, 1998; Rosholm & Skipper, 2009) – as when participation is too low and researchers take various measures to "force" participation, the evaluated programme then becomes far from a "normal" programme (Bernard *et al.*, 2012; Quentin & Guérin, 2013; Morvant-Roux *et al.*, 2014); spillover effects and also attrition effects, which have failed to find a solution despite the many efforts to do so[8] (Eble *et al.*, 2014). These problems long recognised in medicine have since been evidenced in many concrete examples without the randomistas showing any more reflexivity in their research protocols (Shaffer, 2011). So the methodological properties of the RCTs conducted in development are systematically inferior to those conducted in medicine (in North and South alike) and in the developed countries for social policies (Eble *et al.*, 2014).

The question of external validity is by far the most discussed in the literature. The focus on an "average" impact, quite aside from the abovementioned statistical problems, says nothing about the heterogeneity of the impacts and their distribution (Ravallion, 2009; DFID, 2012). The restriction to a short-term impact (for reasons of cost and attrition) often means that midpoint indicators are studied, which can be very different from final outcomes (Eble *et al.*, 2014) if not the reverse, since many project trajectories are not linear (Labrousse, 2010; Woolcock, 2009). Knock-on and general equilibrium effects are ignored despite there being all number of them (Acemoglu, 2010; Ravallion, 2009; Deaton & Cartwright, 2016).[9] The same holds true for the political aspect of programme replication, despite its being a key consideration for scale-up (Bold *et al.*, 2013; Pritchett & Sandefur, 2013; Acemoglu, 2010). Last but not least, the reasons for the impact are disregarded: RCTs might be able to measure and test some intervention impacts and aspects, but they cannot analyse either their *mechanisms* or their underlying *processes*. Overcoming this limitation of the probabilistic theory of causality would call for a "causal model" (Cartwright, 2010), a coherent theory of change (Woolcock, 2013), a structural approach (Acemoglu, 2010) and evaluation of the intervention in context (Ravallion, 2009; Pritchett & Sandefur, 2015). It is no small paradox that RCTs presented by their promoters as the only instruments able to identify a real causal impact ultimately have absolutely nothing to say about the processes at work, i.e. the causes of the impact.

---

[8] See Duflo *et al.* (2007) for an example of studies proposing methods to try and correct attrition effects (and other biases). However, these recommendations are seldom taken up in practice (Eble *et al.*, 2014).
[9] It is interesting to note the paper by Cahuc and Le Barbanchon (2009), which is highly critical of RCTs for ignoring general equilibrium effects in the evaluation of public policies. Cahuc himself went on to whip up a huge controversy in France when he co-authored a virulent attack on all economists, who he called "denialists" to be got rid of (as per the book's title) for not considering their discipline as an experimental science with RCTs as its gold standard (Cahuc & Zylberberg, 2016).

RCTs, whatever their scope of application, sacrifice external validity at the cost of internal validity (Cartwright, 2010). In policymaking, Pritchett and Sandefur (2015) suggest that this trade-off is a mistake. Taking examples from the economics of education (class size effects and gains from schooling) and then microcredit, the two authors suggest that it is much more useful for policy decisions in a given context to look to non-randomised trials conducted in the same context than randomised trials conducted in a different context. More generally, they set out to categorically prove that the claim of external validity for RCT-estimated impacts is necessarily invalid and that the resulting policy recommendations are groundless. Taking a different approach in the form of a political economy angle, O'Laughlin (2015) comes to a similar conclusion in epidemiology, HIV to be more precise, based on the southern African case.

All of these elements make it hard to use RCTs in any way to improve or design policies, which severely limits their scope of application. Once all of these criticisms have been taken on board, as much in terms of internal as external validity, where does this leave RCTs?

Deaton and Cartwright (2016) suggest that RCTs nonetheless remain valid in two areas: 1) to test a theory; and 2) for the ad-hoc evaluation of a particular project or policy in a given context provided that the potential internal validity problems are solved. This restricts their scope to a very narrow spectrum (Picciotto, 2012), dubbed "tunnel-type" programmes by Bernard *et al.* (2012). These programmes are typified by short-term impacts, clearly identified, easily measurable inputs and outputs, and uni-directional (A causes B) linear causal links, and are not subject to the risks of low uptake by targeted populations. They echo the suggestions made by Woolcock (2013) that projects subjected to randomisation need to exhibit low causal density, require low implementation capability and feature predictable outcomes.

Taken together, these conditions rule out a large number of development policies involving combinations of socioeconomic mechanisms and feedback loops (emulation effects, recipient learning effects, programme quality improvement effects, general equilibrium effects, etc.). In the terms of reference for a study commissioned on the subject, a group of DFID managers estimated that less than 5% of development interventions are suitable for RCTs (DFID, 2012). Although this figure is not to be taken literally, there is no doubt that experimental methods are not suitable to evaluate the impacts of the vast majority of development policies. In their more formalised paper, Sandefur and Pritchett (2013) come to a similar conclusion.[10]

Restricting the field of impact evaluations to interventions likely to meet the tenets of randomisation not only rules out a large number of projects, but also many structural aspects of development, both economic and political, such as corporate regulation, taxation and international trade to name but a few. In the face of often complex and systemic causalities, instead of asking "what works", should the question not be rather how to promote flexible, adaptive, creative projects, as suggested by Pritchett *et al.* (2013)? This implies including the

---

[10] "The scope of application of the 'planning with rigorous evidence' approach to development is vanishingly small," Sandefur & Pritchett, 2013, p. 1.

collection of useful data in the projects and then setting up monitoring systems for iterative learning to gradually be able to adjust the project to a complex, volatile environment.

While some of the most prominent promoters of RCTs acknowledge that RCT findings are closely associated with each specific context in which RCTs are used (time, place and project intervention methods), they still argue that they should be considered a "global public good" and an international body created to scale them up (Savedoff *et al.*, 2006; Glennerster, 2012). Such a body would build a universal database and act as a "clearing house", providing answers on what works and what doesn't work in development (Duflo & Kremer, 2005; Banerjee & Hee, 2008). This scale-up plan is illustrated on the J-PAL website, which features eleven evaluated scaled-up projects – police skills training for the "political economy and governance" sub-heading, deworming and remedial education for "education", and free insecticidal bednets for "health" – with 202 million beneficiaries.

RCTs focus on small, relatively simple and easily actionable set-ups, which cannot possibly combine to represent all development issues or form any basis for a social policy. Their above-discussed external validity limitations dispel the randomistas' claim to offer a basket of global policies based on necessarily local RCTs, all the more so since there are no laws in social sciences as there are in the "hard" sciences (physical and natural). This means that there are no universal parameters that can be deemed equivalent to the gravitational constant, Euler's constant, etc. We therefore believe scale-ups (e.g. nationwide) of policies evaluated in experimental conditions and the associated need to rely on structurally weak public institutions to be a particularly thorny political economy issue (see below). François Bourguignon, a prominent researcher who has largely contributed to promoting RCTs, has dubbed this proposal crazy and scientifically impossible.[11] Given these circumstances, pursuing this gargantuan project is at best impetuous, but is more probably driven by interests that need to be identified. We start with those concerning evaluations in the field.

2.2. Scope (in practice): the political economy of four RCTs in the field

Science studies, based on the work of Bruno Latour, show the extent to which the production and use of scientific findings, whatever they may be, are interconnected with the socio-political dynamics around them. Findings cannot escape this political melting pot, not even quantified results that might well be thought incontrovertible and usable as such. A preliminary translation

---

[11] François Bourguignon was the Director of the Paris School of Economics from 2007 to 2013. He served as Chief Economist and Senior Vice President at the World Bank in Washington from 2003 to 2007, during which time he contributed to the creation of DIME. In his closing address to the AFD-EUDN conference in Paris on 26 March 2012, he said (excerpts), "There has been this fashion during the last couple of years on the RCTs. We even heard colleagues, good colleagues, saying that in the field of development, and in the field of development aid, the only fruitful approach from now on was to do random control trials in all possible fields of interventions. And at the end, we'll have a huge map, a huge catalogue saying, 'This works, this doesn't work.' This is crazy! This will never work and, because of that, we absolutely need the other approaches to evaluating policies and programs. 'Pure, scientific evidence' on all that is concerned with development is simply completely impossible. We have to live with this imperfect knowledge." *(our underlining)*

process is needed to put them into words (to make the transition from econometrics to text). Their dissemination and then their reception entails a second translation process: they are subsequently reappropriated, transformed, and sometimes twisted and subverted by different rationales that are hard to predict as they depend on singular historical, social and political contexts. No understanding of the impact of RCTs can dispense with this type of analysis. In other words, RCTs themselves need to be placed under the microscope of a political economy analysis. The following section examines the political economy of the method as a whole. Few empirical elements are available on actual RCT implementation since they would assume an ability to analyse these evaluations' implementation processes, which are often poorly documented and from which external observers and critics are often excluded.[12] We have nonetheless managed to gather together detailed elements on four liberally cited iconic RCTs conducted by the field's most distinguished researchers and published in leading journals. We take these examples to illustrate the political economy of these evaluations, i.e. how these RCTs, rigour (internal or external) aside, are subject to a host of socio-political influences as much in their design and execution as in the dissemination of their results.

Take first the emblematic case of the evaluation by the International Food Policy Research Institute (IFPRI) of the Progresa programme, later called Oportunidades and then Prospera. The programme was set up by the Mexican government in 1997 to provide cash transfers to poor households in return for their compliance with a certain number of requirements designed to improve their children's education and health. The programme was behind the rollout of conditional cash transfer (CCT) policies in developing countries in the late 1990s. This example shows the considerable impact an evaluation can have on public policy (and the method's credibility), despite mediocre internal validity passed over in silence.

Faulkner (2014) meticulously unearths the evaluation's (unpublished) technical documentation, showing that both the initial protocol (choice of sample) and its implementation (attrition and spillover effects) depart substantially from the theoretical framework for RCTs. For example, on the first point, the initial treatment and control group samples were chosen from two different universes rather than being drawn at random as required by the very principle of randomisation. Yet this did not prevent subsequent publications from presenting the sampling protocol as truly experimental. Only by gradually losing sight of these shortcomings was the Progresa evaluation able to be "sold" as the founding example of RCT validity for the estimation of the causal impact of social programmes in developing countries.

As the author points out, emphasising (or even mentioning) the evaluation's weak points would probably have been counterproductive given what was at stake, i.e. the international promotion of RCTs as both an original and the most relevant instrument to evaluate the impact of development programmes and also, in the Mexican case, keeping the programme going after the upcoming presidential elections in 2000 and the impending change in power. It was in the interest of all the players involved in the programme (researchers, promoters and decision-

---

[12] See Jatteau, 2016, pp. 56-66.

makers) for the study to be flawless in its method and convincing in its results (Faulkner, 2014: 239).

Some authors even go so far as to consider that both the experimental design of the evaluation protocol and the assertion that the positive effects measured stem from the conditional nature of the programme were improperly claimed after the fact. They believe this suited political motives to secure the programme's future following the democratic switch in power, given that the evaluation in actual fact showed that the programme was cost-ineffective as a mechanism to raise the school enrolment rate (Pritchett, 2012; Shah *et al.*, 2015).

The second example is drawn from our own observations.[13] It concerns an evaluation of the SKY microhealth insurance programme in Cambodia funded by a French donor (French Agency for Development – AFD) and conducted by a North American team (Center of Evaluation for Global Action – UC Berkeley) in partnership with a local consultancy firm (Domrei). The programme is run by a local NGO with technical assistance from a French NGO (GRET). An "evaluation of the evaluation" was carried out based on an ex-post reconstruction of e-mails, progress reports, meeting minutes and interviews with the study's main protagonists. As with all development projects (the study lasted over five years and cost more than a million dollars from drafting the first terms of reference to results dissemination), the very nature of the study was largely shaped by the complex interplay of stakeholders and their different and sometimes incompatible interests.

The donor, the lead project funder commissioning the study, was interested in testing the feasibility of randomised studies more than the impact evaluation itself (albeit naturally not indifferent to the evaluation, especially when the results were disseminated). The two NGOs, in need of funding, had no choice in the matter and sought essentially to prevent the study from disturbing their day-to-day activities (although they too paid close attention to the content of the results when they were disseminated). Top of the research team's agenda was for the study to contribute to the academic debates on adverse selection in insurance, a subject as yet unexplored in the Southern countries and therefore with huge publication potential. This mixed bag of priorities, inevitable in a project of this scale, weighed on the study throughout. As a source of tension and repeated conflict, it resulted in many compromises as much in terms of research aims, sampling and questionnaire design as in the interpretation and dissemination of the findings. These compromises came down mainly (but not exclusively) on the side of the American academic team, probably due to its prestigious standing (in the French and Cambodian context), but also to the expertise required to make a contribution to the debate on the method. Two lessons can be drawn from the project.

First, here again, close analysis of the survey protocol places a serious question mark over the evaluation's internal validity, a point the final publication fails to mention (Levine *et al.*, 2016).

_____

[13] One of us conducted the study, detailed in Quentin and Guérin (2013), and the other two authors of this article were on the study's steering committee.

Following the ethical and practical demands made by the two NGOs (which refused to disrupt their microinsurance subscription rules and won their case on this score), random sampling was restricted to a lottery draw of volunteers at village meetings. A persistent discrepancy can be observed between the names of the people sampled and the names of the people surveyed, which would seem to suggest some foul play. A very low take-up rate urged the researchers to ask for special information campaigns, fieldworker incentives and discount rates to be put in place, such that the evaluated programme became far from a "normal" programme.

Secondly, and this is the most edifying point, the presentation and dissemination of the results can only be understood in the light of the interplay of stakeholders. Low participation and high dropout rates are among the key conclusions, bearing out the two NGOs' observations since the beginning of the 2000s. As regards the impact itself, the major finding is a significant reduction in household health expenditure and debt, especially among those that have suffered large health shocks. No significant impact is observed on the insurance beneficiaries' health, although there is no way of being conclusive on this point: the study's short timeframe and very low occurrence of large health shocks ruled out any possibility of detecting any statistically significant effects. Impact on the quality of healthcare was quickly dropped from the study as being "unmeasurable" despite being a priority goal for both NGOs and donor. Interestingly enough, the evaluation reports and public presentations made of them, especially to public policymakers in Phnom Penh, highlighted certain positive aspects and failed to mention the more negative findings, in an illustration of the compromises ultimately made by the different players. Most importantly, they did not ask the fundamental question that was the key issue for the NGOs: is voluntary, paid insurance preferable to compulsory or free insurance? In the final academic paper,[14] this question of take-up and the price elasticity of take-up was mentioned, but took a back seat behind the question of impacts.

The third example concerns the problems with upscaling an intervention designed for a town or village to a region or even a country, making nationwide roll-outs of small-scale local programmes problematic. Scaling up implies more than just technical considerations (externalities, spillover effects, saturation, general equilibrium effect, etc.). It is also a question of political economy. A contract (rather than tenured) teacher programme in Kenya provides a typical example of this. An RCT of the programme piloted on a small scale by an NGO had returned a positive impact on pupils' test scores (Duflo *et al.*, 2012). These positive findings appeared to be especially robust in that they supported other results obtained for a similar programme in India (Muralidharan & Sundararaman, 2011).

However, Bold *et al.* (2013) showed that the programme's effects disappeared when rolled out nationwide and government implemented. The study by Bold and her co-authors, based on a highly original RCT embedded with the programme's national scale-up (the first of its kind to test organisational and political economy effects), concludes that the absence of effects following

---

[14] The article was co-authored with the local consultant, who was highly critical of the validity of the method and its implementation, which is probably another illustration of the compromises that ended up being made.

upscaling was due to the change of project operator: from carefully selected, highly motivated NGOs to unionised government workers. This bias could even be systematic where there is a correlation between the people, places and organisations that opt into implementing the RCTs and the estimated impacts (Pritchett & Sandefur, 2013). We believe that Acemoglu (2010) has a particularly decisive theoretical argument to make when he discusses political economy responses to large-scale programmes from groups who see their rents threatened by reforms. Vivalt (2016) shows that this is more than just a theoretical argument or specific to the case presented here: one of the main findings of her meta-analysis is that government programmes have significantly weaker effects than programmes implemented by NGOs and research centres. This example places two limitations on replicating RCT results for locally conducted policies considered for scale-up. Firstly, the type of operator conducting the policy is decisive (here, public institutions versus NGOs). Secondly, the proliferation of RCTs conducted in different fields (here, India and Kenya) in no way makes results more robust, as the decisive factor for the programme's success is precisely the type of operator conducting it.

Note that even in the article published a good three years after the controversy sparked by the Kenyan experiment, Duflo and her co-authors (2015) made but marginal amendments to their conclusions. Bold *et al.*'s work (which has still not been published despite its major interest) only gets a mention in passing and in the conclusion even though the authors question both the external validity of RCTs and, implicitly, their internal validity: the treatment players – the teachers – are aware of their status and manipulate the experiment.

Our fourth example is the evaluation of the deworming programme for Kenyan children in the late 1990s with its massive impact still today. Two economists from Berkeley and Harvard, Miguel and Kremer (2004) concluded in a famous article published by the prestigious *Econometrica* journal that deworming had a significant impact on health and school attendance, but not on academic test scores. Although the authors identify a direct impact on the treated children, their emphasis is above all on the indirect impact: a positive externality effect reducing contamination among untreated children on contact with the treated children. This is the mechanism pointed up in the article's title and is probably the reason for its publication in *Econometrica*. They conclude from their cost-benefit analysis that this is the most effective way to improve school participation.

This study is one of the flagships of the RCT movement. It is the direct inspiration for the Deworm the World campaign, which raised 2.5 million dollars in 2013 and 15.8 million dollars in 2014 (latest figures available). J-PAL holds up this founding RCT as the subject of the largest number of scale-ups (5 in 11) and the one that has reached the most people: 95 million of the reported 202 million (Duflo, 2016).

This study, long seen as the RCT movement's star success story, has recently found itself at the centre of a controversy that has caused a media stir worldwide (Boseley, 2015). In a replication programme funded by 3ie, epidemiologists from the prestigious LSHTM laboratory disputed the initial study's findings. In a first paper, they reanalysed the micro-data and demonstrated that a

certain number of results did not hold (Aiken *et al.*, 2015). They identified numerous discrepancies in the handling of the data, including a massive missing data problem. Once these errors were dealt with, the total effect (direct and indirect) on the reduction in school absenteeism was half that initially stated and was no longer significant: although the direct effect on treated children continued to hold, there was no positive externality on neighbouring untreated schools as the initial study's signature finding had it.

In a second paper, Davey and his co-authors (2015) re-analysed the data using what the epidemiologists judged the most appropriate statistical impact estimation methods to analyse the specific case of the randomisation protocol actually used by the project. They concluded that the project definitely had an impact on school absenteeism, but that the impact was weak and probably biased. In addition, it was not possible to ascribe this effect exclusively to deworming, since the other project component (health education) might have been the sole contributory factor. This reservation is all the more plausible in that the project had no impact on the health indicators (weight for age and height for age), midpoint outcomes through which deworming would logically have had to have channelled to have an effect on the children's absenteeism. The epidemiologists more generally criticised the failure of the randomisation protocol and subsequent analysis to subscribe to the codified scientific standards required in the field of public health RCTs.

Without going into how this controversy has subsequently developed (see Humphrey, 2015), we conclude from this example not only that the prerequisites for programme scale-up were not met (external validity), but also that the initial study's internal validity was itself not guaranteed. To crown it all, the randomistas in this controversy have been taken to task by other randomistas (epidemiologists), the former of whom claim direct methodological descent from the latter, in a sort of evidence-based medicine versus evidence-based policy dispute. All of this raises the question as to which factors could have driven the programme's scale-up in what was an especially risky rollout given the fragility of the results obtained. This is precisely the question addressed by the last section of this article taking a broader political economy angle.

III.   Political economy of a scientific enterprise

Any understanding of the contradictions between the method's limitations and its huge credibility, in both the academic and political field, first needs to consider the balances of power at work that go into shaping collective preferences for one or another method. Impact evaluations, with RCTs as their ideal model, have in this way become so massively widespread that they have turned into quite the industry. As with any industry, the impact evaluation market is where supply meets demand. Demand is twin-engined, driven by both the donor community and the academic world. Supply is largely shaped by a brand of scientific businesses and entrepreneurs, which we undertake to describe in this section along with the strategies they use to "corner" the market.

3.1. A new scientific business model

Looking first at the donors, the second half of the 1990s and the 2000s saw the "end of the ideologies" so characteristic of the structural adjustment era. With the end of the Cold War, the political sphere started to ease its grip on official development assistance (ODA). Technical and financial cooperation during the Cold War was often merely another pawn in bloc rivalry. As the Berlin Wall fell, so too did cooperation's subordination to realpolitik. In the new post-modernist world, ODA promoters have found themselves under the spotlight as the aid crisis, MDGs and New Public Management have summoned them to the stand to prove their utility (Naudet, 2006).

The new credo focuses development policy on poverty reduction and promotes results-based management. These guidelines were formulated in the 2005 Paris Declaration on Aid Effectiveness and thereafter systematically reiterated by the major international conferences on official development assistance in Accra in 2008, Busan in 2011 and Addis Ababa in 2015. The rise of the evidence-based policy paradigm, which consists of basing all public decisions on scientific evidence, has given scientists new credibility in these political arenas. RCTs in principle meet all the conditions required by this game change: agnostic empiricism, apparent simplicity (simple comparison of averages), elegant use of mathematical theory (guarantee of scientificity) and focus on the poor (household surveys). Their simplicity makes them easy for policymakers to understand, lending them appeal as a vehicle for informing public decision-making. The evaluation of the Progresa programme in Mexico formed a prototype for this method and a textbook example of its performance capabilities, as discussed above.

The academic climate, especially in economics, is also conducive to the rise of RCTs: demise of the heterodox schools concentrating on social structures and domination processes, search for the microfoundations of macroeconomics, primacy of quantification and economics in the social sciences, and alignment with the standards holding sway in the North (Milonakis & Fine, 2009). The joint rise of behavioural and experimental economics, capped by the 2002 award of the Nobel Prize in Economics to psychologist Daniel Kahneman and economist Vernon Smith, respective experts in each field, shows just how far the discipline has come. Yet RCTs are fuelled by and in return fuel this rise, which is itself a subject of heated debate.[15] Experimental economics is a way of producing controlled and replicable data, with different variants depending on the level of control from laboratory trials through to natural experiments. RCTs provide the opportunity to conduct experiments *in the field*, thereby extending their scope to various purposes that do not (or inadequately) lend themselves to laboratory experiments (List & Metcalf, 2014). Behavioural economics, largely but not solely based on experimentation, is defined by its purpose: analysis of cognitive, emotional and social biases in individual behaviour. Although it criticises the descriptive aspect of neoclassical economics' hypothesis of rationality, it retains its normative dimension in the form of the recommendation of tools – generally *nudges* – supposed to correct behavioural imperfections. RCTs draw extensively on the precepts of behavioural

---

[15] See Teele (2014) for an overview of the controversies in field experiments and Kosters *et al.* (2015) for behavioural economics.

economics,[16] and have actually been the vehicle that has channelled behavioural economics into development economics to the extent that it now occupies a dominant position in the discipline (Fine & Santos, 2016).

The World Bank, a major player with dual credibility as both financial and academic donor, has also been a catalyst in the rise of both the evidence-based policy paradigm and RCTs. First of all, it was the scene of a scientific U-turn away from classical (macro)economic development studies, the bastion of which was the World Bank's research department, towards new empirical approaches with a microeconomic focus. The seeds of this turnaround were sown in 2003 when François Bourguignon was appointed Chief Economist. In 2005, he contributed to the creation of a dedicated impact evaluation unit (DIME), financed by research department funds. He also commissioned an evaluation of the research department's work. This evaluation lambasted the scientific research conducted by the Bank in the previous decade for being essentially, "used to proselytize on behalf of Bank policy, often without taking a balanced view of the evidence, and without expressing appropriate scepticism [*and*] a serious failure of the checks and balances that should separate advocacy and research," (Banerjee *et al.*, 2006, p. 6).

This criticism was echoed in a report by the international Evaluation Gap Working Group comprising many renowned researchers, including the foremost advocates of RCTs (F. Bourguignon, A. Banerjee, E. Duflo, D. Levine, etc.), and leading development institution heads (DAC, World Bank, Bill & Melinda Gates Foundation, African Development Bank, Inter-American Development Bank, etc.). *When Will We Ever Learn?*, published by the Center for Global Development (Savedoff *et al.*, 2006) in the form of a call-programme, was taken up far and wide by the scientific community, practitioners and policymakers. In addition to its arguments, the report also acted as self-serving advocacy since it raised the profile of and demand for studies from many of its authors, first and foremost RCTs.

The 2015 World Development Report marks the most accomplished convergence of the two movements to date: in support of RCTs as a methodology in general and in support of behavioural economics as a disciplinary approach. It sets out to redesign development policies based on a "richer view of human behaviour" and the use of nudges (defined in the Report as, "A policy that achieves behaviour change without actually changing the set of choices,"; World Bank, 2015, p. 36) to correct behavioural imperfections. This move is representative of the abovementioned entanglement of RCTs, behavioural economics and experimental economics (Fine & Santos, 2016).

Yet the pro-RCT momentum has been driven above all by the emergence of a new generation of researchers. They are young and from the inner sanctum of the top universities (mostly American).[17] They have found the formula for the magic quadrilateral by combining the mutually

---

[16] The work of Banerjee and Duflo (2011) and Karlan and Appel (2012) are highly illustrative of this.

[17] A. Jatteau (2016) shows that J-PAL researchers are both more often graduates from elite establishments and hold more prestigious positions than their counterparts on average (chairs, Ivy League Plus, NBER, BREAD, CEPR, etc.).

reinforcing qualities of academic excellence (scientific credibility), public appeal (media visibility and public credibility), donor appeal (solvent demand), massive investment in training (skilled supply) and a high-performance business model (financial profitability). With a multitude of university courses and short training sessions for a wide audience taught in classic (face to face) and new forms (MOOC), the randomistas have devised the means to attract young, motivated and highly skilled resources.[18] In an intense whirl of communication and advocacy, backed by a plethora of press and para-academic media (policy briefcases, blogs, outreach forums, happenings, etc.), they give the welcome impression of researchers stepping out from their ivory tower. Their modest, grassroots position[19] embodies commitment, empathy and impartiality.

A study of the career paths of high flyers and their networks can be a compelling angle from which to understand the emergence, decline and transnational spread of scientific and policy paradigms (Dezalay & Garth, 2002). It is not our intention here to study this angle in full with respect to RCTs, which would form a research programme of its own. We will settle instead for an outline. The analysis by A. Jatteau (2016, pp. 305-334) on randomistas' networks and the connections between them based on the body of researchers involved in conducting RCTs and authors of associated academic publications finds evidence of a dense, hierarchical network of interrelations from which two leaders emerge ("nodes" in network theory): Dean Karlan and Esther Duflo, the movement's figurehead. This young French-American researcher has a string of academic distinctions to her name, including the distinguished Bates Medal for the "best economist" under the age of forty in 2010. She has an impressive number of publications to her credit in the most prestigious economic journals. Yet she also makes her work more widely available in the form of publications for the layman and bestsellers (see, for example, Banerjee & Duflo (2011) in English and Duflo (2010) in French). US magazine *Foreign Policy* has consistently included her on its list of top 100 global intellectuals since 2008. In 2011, *Time Magazine* named her one of the 100 most influential people in the world. In late 2012, she was appointed advisor to President Obama on global development policy. In France, she was the first to hold the brand new Collège de France "Knowledge Against Poverty" Chair, funded by AFD (French Agency for Development). In 2015, she was presented with the Princess of Asturias Award for Social Sciences, a prestigious Spanish distinction. Her name regularly comes up as potentially in the running for a future Nobel Prize in Economics.

These young RCT movement researchers have also made a name for themselves with their management methods. By setting up NGOs and specialised consulting firms, they have created suitable structures to receive funds from all sources: public, naturally, but also foundations, businesses, patrons, and so on that are off the beaten public research funding track. From this point of view, they are in perfect harmony with the new sources of aid financing from private foundations and philanthropic institutions, which are particularly inclined to entrust them with

---

[18] As shown by the successful dedicated online courses on MIT's edX platform; see, for example, *The Challenges of Global Poverty* MOOC taught by Banerjee and Duflo and the *Evaluating Social Programs* MOOC taught by two of their J-PAL colleagues and taken by over 1,000 participants in 2013.

[19] As mentioned before, the sheer number of RCTs ongoing at the same time places a question mark over their real knowledge of the ground.

their studies. By managing to create their own funding windows – mainly multilateral (World Bank initiative for the development impact evaluation, international impact evaluation initiative, African Development Bank and Strategic Impact Evaluation Fund), but also bilateral (Spanish and UK cooperation agencies) and from major foundations (Rockefeller, Citi, Gates, MacArthur and Hewlett) – the randomistas have created an oligopoly on the flourishing RCT market, despite keener competition today due to the adoption of RCT methods by a growing number of research teams.

The loose conglomeration that has formed around J-PAL is the most emblematic and accomplished example of this new scientific business model. The J-PAL laboratory is attached to the MIT Economics Department. These institutional roots, with one of the top American universities, and the high profile of its directors act as both an academic guarantee and a catalyst.

Innovations for Poverty Action (IPA) plays a key role alongside J-PAL. This non-profit organisation has managed to generate over 250 million dollars in revenue since 2003 when it was set up, posting steadily growing sums every year (Jatteau, 2016, p. 265). In addition to its RCT communication and advocacy role,[20] it works to scale up and replicate randomised control trials once they have been tested by J-PAL. The combination of the two institutions therefore has the set-up to accomplish the scale-up plan described in the second section. Annie Duflo, Esther Duflo's sister, is IPA's Executive Director. Dean Karlan (mentioned earlier for his 100 RCTs), Professor at Yale who studied for his PhD under the two J-PAL initiators, is founder and board member. And with Abijit Banerjee also being Esther Duflo's life partner, J-PAL/IPA is more than a global enterprise; it is also a family affair. More broadly speaking, the borders between the two institutions are porous and many members and associates have cross-cutting responsibilities in both.

The RCT industry is a lucrative business in every respect. It is academically rewarding, and there is everything to be gained from joining this movement (or everything to be lost from not being in it). Today, it is very hard to publish papers based on other approaches in the economic journals. This crowding-out effect also ties in with the fact that the most influential RCT promoters are often on the editorial boards of the leading economics and development economics journals.[21] The *American Economic Journal: Applied Economics'* special issue on RCTs of microcredit is illustrative in this regard. The issue's three scientific editors are members of J-PAL. In addition to the general introduction, each editor co-signed a paper and two of them were members of the

---

[20] Unlike J-PAL, which needs to comply with academic decorum, IPA can state its aims in marketing terms. For example, "IPA uses randomized evaluations because they provide the highest quality and most reliable answers to what works and what does not," (see the IPA website: http://www.poverty-action.org/).

[21] Among the most well known being the *Annual Review of Economics, Journal of Economic Literature, the American Economic Journal: Applied Economics, Review of Economics and Statistics, Journal of Development Economics, Review of Development Studies, Journal of Quantitative Economics* and *Journal of Economic Perspectives*. Here again, D. Karlan turns up as the most voracious player: he is on at least eight boards, including five of the abovementioned publications, along with *Behavioral Science & Policy, Stanford Social Innovation Review* and *Journal of Globalization and Development*.

board of editors (Banerjee and Karlan). Esther Duflo is both the journal's editor (founder) and co-author of two of the six papers. Given in addition that nearly half of the papers' authors (11 of the 25) are also members of J-PAL and four others are affiliated professors or PhD students with J-PAL, the journal has strayed somewhat from the peer review principles supposed to govern scientific publication. This single example shows in cameo the extraordinary density of the links between randomistas identified by Jatteau (2016).

Yet the rewards are more than just symbolic. Specialising in RCTs is also an excellent way to find a position as a researcher or teacher, as shown by current recruitment methods in economics. And it guarantees the securing of substantial funds to conduct own research (at a time when funds are in short supply everywhere) and huge additional earnings from consultancy and sitting on management bodies (Jatteau, 2016).

3.2. Market control and rent capture strategies

Given these circumstances, it is easier to understand why criticism of RCTs is greeted with hostility and fiercely opposed by RCT promoters. A number of strategies are employed to underpin the monopoly and supremacy of RCTs. As was the case in medicine, alternative methods are discredited as RCTs assume the scientific monopoly (Harrison, 2011). Reference to evidence-based medicine is put forward as a guarantee of scientific integrity, but without mention of the many controversies it has sparked and which are still ongoing today (Jacobson *et al.*, 1997; Schulz *et al.* 1995; Labrousse, 2010). Face-to-face debate is often sidestepped or refused.[22] Critical voices have long remained out of earshot, confined to the pages of marginalised publications. In many cases, the results of trials presented as new "discoveries" are really nothing more than rehashes of conclusions from past studies. The subterfuge is a two-step process. Firstly, most of the existing literature is discredited as not being rigorous on the pretext that RCTs are superior and considered the only conclusive way of producing evidence. This virtually ritual denigration effectively wipes clean the memory banks of past knowledge. The resulting reset effect means that all RCT findings can be passed off as major "discoveries" despite their being potentially (and often) redundant.[23] The ploy is particularly plain to see in that published papers based on non-experimental methods are virtually never cited (Labrousse, 2010; Nubukpo, 2012).

So the randomistas' takeover is not just a question of method. It extends to content and the arguments that hold currency on the major scientific and political issues in the development field.

---

[22] See, on this subject, A. Deaton's interview with A. Jatteau (2016, p. 60). It is enlightening to take a look at A. Jatteau's scientific career itself as described in his dissertation, whereby he was distanced for his critical designs from the J-PAL network of which he had been a member. On our more modest level, all attempts to engage in a debate with the movement's promoters (via publications or conferences) to hear both sides of the story have been evaded.

[23] This well-known modus operandi has already been used in development economics with respect to the institutions. Having knocked the schools of heterodox economics (regulation school, convention school, neo-institutionalism, etc.) off the board of legitimate disciplines, despite development economics being one of their originalities, mainstream economics re-appropriated the subject and passed its findings off as new. The same phenomenon is currently taking hold in political economy.

In addition to the abovementioned iconic deworming case, a few microfinance examples can be taken to provide a good illustration of this tendency to pass off as original theories already largely explored in the literature (Bédécarrats *et al.*, 2015).[24]

The difference between literature and practice paints a clear picture of the relative independence of mainstream research from practice and its imperviousness to otherwise-robust non-mainstream studies. Academic literature published in leading international journals became infatuated with the so-called financial revolution of group lending, even though it had been practised by many microfinance institutions since the 1960s (Gentil, 1996; Dichter & Harper, 2007). The question of the effectiveness of group lending was eventually settled in a most down-to-earth manner by field players who developed individual lending (starting with the Grameen Bank) in the mid-1990s. In the early 2000s, a branch of more operational research showed that group lending is only suitable in the right social and economic circumstances (dense, but relatively flat social networks, a low level of borrower specialisation in the same sector, etc.). Other academic studies and research based on robust quantitative studies come to similar conclusions (Godquin, 2004; Sharma & Zeller, 1997; Gonzalez-Vega *et al.*, 1996). Yet experimental studies totally ignore these different branches of literature and contribute nothing new, aside from the way they produce the evidence. For example, the studies by Giné and Karlan (2014) of the effects on loan default of group (versus individual) liability lending are superfluous considering the dozens of studies already conducted on the subject, especially the three cited above. More broadly speaking, there is really nothing new about the randomistas' conclusion that, "Microcredit therefore may not be the "miracle" that is sometimes claimed on its behalf, but it does allow households to borrow, invest, and create and expand businesses," (Banerjee *et al.*, 2015a). It is found in many of the 170 impact studies (including a dozen RCTs) published on microfinance from 1980 to 2010 (Bédécarrats, 2012).

There are many more examples to be found. Whether in group (versus individual liability) lending or savings, pre-RCT literature prompts the question as to whether RCT conclusions are really new. They come more than 15 years after the financial innovations were tested, documented and rolled out. This observation contradicts the message aired by J-PAL and IPA, who present themselves essentially as test pads to stimulate social policy innovations.[25]

Despite the gathering of clouds for an outright scientific controversy (as defined by Callon *et al.*, 2001; Knorr-Cetina, 1982) over RCTs, the power imbalance between stakeholders has so far prevented any full-blown storm. Note, however, that criticism is growing and slowly redrawing the lines as shown by the turning tide in the scientific field detailed by our series of studies on the issue (Bédécarrats *et al.*, 2013, 2015 and this article). Not all of the otherwise minority critics have the same reach or voice. It is harder, for example, to shrug off Angus Deaton's persistent attacks on the randomistas' hegemonic plans and, following his 2010 article, his new harshly critical

---

[24] The theories developed by Banerjee and Duflo (2011) in their book entitled *Poor Economics* are alone worth putting under the microscope from this point of view.
[25] See IPA's description of its strategy on its webpage: http://www.poverty-action.org/about, consulted on 22 January 2015.

article co-written with Nancy Cartwright, internationally renowned philosopher of science and expert on RCTs in medicine, especially as he has since been awarded the Nobel Prize.[26]

Here again, the special issue on microcredit could be mentioned by way of an example. The issue was published with the original databases in response to the complaint about opaqueness and to facilitate meta-analyses. The general introduction summarises the responses provided (Banerjee *et al.*, 2015a). A theoretical model is developed in response to the agnostic empiricism criticism. The issues of take-up rate, estimator accuracy and treatment heterogeneity are acknowledged (internal validity). Contextual diversity is addressed by the range of settings, products and institutions covered by the six papers (external validity). The cost-benefit analysis proposed by Banerjee *et al.* (2015a) is supposed to answer the question of efficacy. Yet the shift in position remains but slight. Basically, the two main founding principles remain: i) RCTs are deemed the most rigorous way of measuring causal impact, and ii) the scale-up plan designed to establish what works and what doesn't work stands unchanged. This turns the argument about factoring in the diversity of contexts on its head. The similarity of results obtained in the six countries considered to be "fairly representative of the microcredit industry/movement worldwide" (Banerjee *et al.*, 2015a, p. 2) allegedly resolves the criticism of the method's external validity.

This sidestepping of the controversy is especially problematic considering that the stated end objective of RCTs – to inform public policy – is largely subject to caution. Eleven, or 2%, of the 543 RCTs conducted and completed by J-PAL (of the 811 launched) have actually been scaled up. Note also that five of these concern deworming, whose flaws are covered in Section II, and three others cover the distribution of chlorine, whose contribution has been seriously nuanced by LSHTM (Wolf *et al.*, 2014). This point really merits a fully-fledged analysis, but three arguments can nonetheless be put forward. The first is the complexity of policymaking, which also explains why experimental methods, irrespective of their scope, actually have a limited direct effect.[27] The second argument, already raised, concerns the narrow focus of the questions likely to comply with the tenets of randomisation, which rules out a vast swath of development questions (see above, IIIa). When the World Bank (2015) sets out to redefine development policies, the proposals are such as metal piggy banks to increase poor households' savings, television commercials to reduce water consumption, stickers in buses to reduce road accidents, and the delivery of fertilisers at the right time to improve small farmers' living conditions. Any taking up of the idea that RCTs could have an impact on development policies calls for the acknowledgement that this then comes down to helping populations make better choices, but in what remains an unchanged environment.

The third argument is associated with the objectives and constraints of academic publication, dictated by the considerable publishing pressure placed on the profession. This weighs heavily on the implementation of RCTs, as shown by Arthur Jatteau (2016) in a thorough analysis of J-PAL's production and numerous interviews with randomistas and their research assistants.

---

[26] Without going into an analysis of the 2016 article, note that it develops an entire battery of new criticism to the extent that the new paper does not even make reference to the 2010 paper.

[27] See, for example, Moffitt (2014) on Social Welfare Programs.

Whether in their choice of subject matter, sampling, questionnaire modules or the type of results put forward, the choices made are often to optimise the ratio of human and financial resources invested to the number of publishable units the work can yield. Academic journals are interested primarily in innovative projects (whatever they may be) and not in the consolidation of already published knowledge in new contexts. These trade-offs are often made at the expense of the expectations of the trial's operational partners, relevance for the populations and utility for policymaking. Many RCTs are out of touch with local realities (Barrett & Carter, 2010), but sell academically (Jatteau, 2016, pp. 423-432) with two particularly frequent biases: too focused on new, supposedly innovative projects and on positive results (when negative or zero effects would be just as instructive, but often go unmentioned). The contradiction between the priority placed on academic recognition by the most prominent randomistas and the ambition to operationally contribute to public policy, which the association between J-PAL and IPA seeks to resolve, could ultimately bring everything grinding to a halt as the RCT funders' upscaling hopes, themselves nurtured by the randomistas, are dashed.

Conclusion

This paper sets out to describe the meteoric rise of randomised control trials in development to the point where they have become the gold standard for impact evaluations. The article goes on to show the methodological limitations of RCTs and the vanity of the hegemonic plans entertained by their advocates. Lastly, it takes a political economy angle to understand the factors behind the establishment of this new international standard. We believe that the use of randomised control trials in development is a methodological advance. Yet this small step forward has come with two steps back: epistemological since RCT disciples share a now-outmoded positivist conception of science, and political in terms of the imperialistic nature of an approach that purports to be able to use this instrument to understand all development mechanisms. In addition, it is no small paradox of the randomistas' success story that they have managed to label the method as the only technique able to rigorously identify causal impacts, when RCTs actually only provide evidence of effectiveness and say nothing about the causal mechanisms at work.

Among the possible extensions of this research, two lines of inquiry appear to be promising: one analytical and the other methodological. On the first front, our political economy approach is worth rounding out with historical and science studies research. To take a "Latourian" slant, research on the interactions between scientific output and social conditions, the personality of the actors and even institutional architecture could be usefully applied to the RCT industry, its guardians and its most prominent research centres: interest in *laboratory life* should not be the sole reserve of the "hard" sciences (Latour, 1999). We could also consider historical approaches to randomistas' career paths, as is already the case for captains of industry, politicians and high-profile scientists. The analyses of the J-PAL laboratory by Jatteau (2016), although far from exhausting the subject, show the wealth of information to be gained from this approach.

On the second front, our purpose is not to reject RCTs, since they constitute a promising method ... among others. However, they still need to be conducted by the book and aligned with best practices established in the medical world. Although RCTs are probably fit and proper for certain precisely defined policies, other methods can and should be used. These methods take a pragmatic approach, defining the research questions and methodological tools required on a case-by-case basis with the partners concerned (field operators, donors, etc.). They also draw on a range of methodologies, based on interdisciplinarity, and acknowledge the different ways of producing evidence (statistical inference/comprehensive analysis). The idea is not to reject formalism and modelling, but to make controlled use of them. Moreover, these approaches do not set out to lay down universal laws, but to explain causal links specific to a particular time and place. Qualitative methods are used to contextualise development policies, develop original hypotheses, identify new and unexpected phenomena, and analyse them from every angle, studying the complexity of the causal links and the many, dynamic and contradictory interactions between different entities in a location-specific way. A particularly interesting idea is the iterative learning suggested by Pritchett *et al.* (2013), which calls for a dynamic study design based on ongoing interactions between the results obtained and the project's modalities.

The extent of method interfacing and integration (quantitative, qualitative and participatory) can vary immensely. Rather than systematically relying on the creation of new data, these alternative methods draw on existing data, where appropriate, taken from official statistics and data produced by local development project/policy partners. This approach cuts costs and streamlines hefty survey protocols with their abovementioned potentially negative effects on research quality. It also improves the evaluated programmes' information systems and the statistical apparatus in the countries in which these programmes are located. Where some donors such as AFD (2013) and, less whole-heartedly, DFID (2012) originally jumped on the band wagon, they are now rather more circumspect about the real scope of application of RCTs and their capacity to answer the questions they ask. Let's hope that this return to a less black-and-white position sees some concrete measures in support of alternative and/or complementary evaluation methods.

R. Picciotto (2012) was already asking when the RCT bubble would burst back in 2012. The criticism is growing and the award of the Nobel Prize to A. Deaton, a fervent critic of RCTs, will probably accelerate the process. Clearly, the RCT movement will do everything possible to avoid demotion so that it can continue to benefit from the returns on its dominant position in the impact evaluation field. The randomistas may end up conceding that they were mistaken in their pretensions to want to define what works and what doesn't work in all areas. Yet this turnaround would only happen under pressure and after reaping in the maximum returns. [28] Credible alternatives would also be needed for the bubble to burst. This means actively pursuing thinking and action on methodological innovation.

---

[28] The question could also be asked as to whether the movement's members are sincere in their defence of their project against the critics. Some probably continue to believe, while others aware of its limitations more cynically hope to make the most of the privileges and advantages it still brings. The interviews conducted by Jeatteau (2016) display a range of such motives.

References

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

AFD (2013), *Stratégie 2013-2016 en matière d'évaluations d'impact*. Paris: AFD.

Aiken A. M., C. Davey, J.R. Hargreaves and R.J. Hayes (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.

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

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

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

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

Banerjee A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert and C. Udry (2015b), "A multifaceted program causes lasting progress for the very poor: Evidence from six countries" *Science* 348 (6236): 1-16. [DOI: 10.1126/science.1260799]

Banerjee A. and R. He (2008), "Making Aid Work" in: W. Easterly, *Reinventing Foreign Aid*. Boston: MIT Press.

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

World Bank (2015), *World Development Report 2015: Mind, Society, and Behavior*. Washington D.C.: World Bank.

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

Bédécarrats F., I. Guérin and F. Roubaud (2015), "The gold standard for randomized evaluations: from discussion of method to political economy", *DIAL,* Working Paper 2015-01.

Bédécarrats F., I. Guérin and F. Roubaud (2013), « L'étalon-or des évaluations randomisées : du discours de la méthode à l'économie politique » *Sociologie pratique* 2013/2 (27): 107-122.

Bédécarrats F. (2012), « L'impact de la microfinance : un enjeu politique au prisme de ses controverses scientifiques » *Mondes en développement* 158: 127-142.

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

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

Boseley S. (2015), "New research debunks merits of global deworming programmes", *The Guardian*, 23 July.

Cahuc P. and T. Le Barbanchon (2009), "Labor Market Policy Evaluation in Equilibrium: Some Lessons of the Job Search and Matching Model", *CNRS/Ecole Poytechnique*, Working Paper 2008-40.

Cahuc P. and A. Zylberberg (2016), *Le négationnisme économique et comment s'en débarrasser*, Paris: Flammarion.

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

Cartwright N. (2010), "What are randomised controlled trials good for? " *Philosophical studies* 147 (1): 59-70.

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

Davey C. (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. and N. Cartwright (2016), "Understanding and Misunderstanding Randomized Controlled Trials", *NBER,* Working Paper 22595.

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

Dezalay Y. and B.G. Garth (2002), *The Internationalization of Palace Wars. Lawyers, Economists, and the Contest to Transform Latin American States*, Chicago: Series in Law and Society.

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

Dichter T. and M. Harper (Ed.) (2007), *What's Wrong with Microfinance?*, Rugby: Practical Action Publishing.

Duflo E. (2016), "Randomized Controlled Trials, Development Economics and Policy Making in Developing Countries", World Bank Conference: The State of Economics, *The State of the World*, Washington D.C., online at: https://olc.worldbank.org/content/state-economics-influence-randomized-controlled-trials-development-economics-research-and.

Duflo E. (2010), *Lutter contre la pauvreté*, Paris: Seuil, collection La République des idées.

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 Economic* 123: 92-110.

Duflo E., P. Dupas and M. Kremer (2012), "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools", *NBER*, Working Paper 17939.

Duflo E. and M. Kremer (2005), "Use of randomization in the evaluation of development effectiveness", in: G.K. Pitma, O.N. Feinstein and G.K. Ingram (Ed.), *Evaluating Development Effectiveness.* World Bank Series on Evaluation and Development (7), New Brunswick: Transaction Publishers: 205-231.

Duflo E., R. Glennerster and M. Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit", in: T.P. Schultz and J. A. Strauss (Ed.), *Handbook of Development Economics*, Vol. 4, London: Elsevier: 3895–3962.

Easterly W. (2007), *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Oxford: Oxford University Press.

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

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

Fine B. and A. Santos (2016), "Nudging or Fudging: The World Development Report 2015" *Development and Change* 47 (4): 640-663.

Gentil D. (1996), "Les Avatars du modèle Grameen Bank" *Revue Tiers-Monde* 37 (145): 115-133.

Giné X., D.S. Karlan (2014), "Group versus individual liability: Short and long term evidence from Philippine microcredit lending groups" *Journal of Development Economics* 107: 65-83.

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

Godquin M. (2004), "Microfinance Repayment Performance in Bangladesh: How to Improve the Allocation of Loans by MFIs" *World Development* 32 (11):1909-1926.

Gonzalez-Vega C., R.L. Meyer, M. Schreiner, J. Rodriguez-Meza, G.F. Monje (1996), *Microfinance Market Niches and Client Profiles in Bolivia*, Columbus: Ohio State University Press.

Harrison G.W. (2011), "Randomisation and Its Discontents" *Journal of African Economics* 20 (4): 626-652.

Heckman J.J. (1991), "Randomization and Social Policy Evaluation", NBER Technical Working Paper 107.

Heckman J.J., J. Smith and C. Taber (1998), "Accounting for dropouts in evaluations of social programs" *Review of Economics and Statistics* 80: 1-14.

Heckman J. J., E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation" *Econometrica* 73 (3): 669-738.

Humphreys M. (2015), "What has been learned from the deworming replications: a nonpartisan view", Columbia: Columbia University, online at: www.columbia.edu/~mh2245/w/worms.html.

IEG (2012), *World Bank Group Impact Evaluation. Relevance and Effectiveness*, Washington D.C.: World Bank.

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

Jacobson L., A. Edwards, S. Granier, C. Butler (1997), "Evidence-based Medicine and General Practice" *British Journal of General Practice* 47: 449-452.

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

Jerven M. (2015), *Africa: Why Economists Get It Wrong, African Arguments*, Zed Books, London.

Jerven, M. (2013), *Poor numbers: how we are misled by African development statistics and what to do about it.* Ithaca and London: Cornell University Press.

Karlan D. (2015), *The Multi-Lateral Development Banks: A Strategy for Generating Increased Return on Investment*, Washington D.C.: US Congress, Testimony before the US House Committee on Financial Services, online at www.poverty-action.org/sites/default/files/Dean%20Karlan_ Testimony.pdf

Karlan D.S. and J. Appel (2012), *More than Good Intentions; Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy*, New York: Dutton Press.

Kosters M. and J. Van der Heijden (2015), "From mechanism to virtue: Evaluating Nudge theory" *Evaluation* 21(3): 276-291.

Labrousse A. (2010), "Nouvelle économie du développement et essais cliniques randomisés : une mise en perspective d'un outil de preuve et de gouvernement" *Revue de la régulation* 7 (2): 2-32.

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

Levine R, R. Polimeni and I. Ramage (2016), "Insuring health or insuring wealth? An experimental evaluation of health insurance in rural Cambodia" *Journal of Development Economics*, 119: 1-15.

List J.A. and R Metcalfe (2014), "Field experiments in the developed world: an introduction" *Oxford Review of Economic Policy* 30: 585-596.

McKenzie D. (2016), "Have RCTs taken over development economics?" *World Bank Blog on Impact Evaluations*, online at: http://blogs.worldbank.org/impactevaluations/have-rcts-taken-over-development-economics.

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

Milonakis D. and B. Fine (2009), *From political economy to economics*, London: Routledge.

Miranda J., S. Sabet S. and A.N. Brown (2016), "Is impact evaluation still on the rise?" *Evidence Matters: Improving development and practice*, online at: http://blogs.3ieimpact.org/is-impact-evaluation-still-on-the-rise/.

Moffitt R. A. (2004), "The Role of Randomized Field Trials in Social Science Research A Perspective from Evaluations of Reforms of Social Welfare Programs" *American Behavioral Scientist* 47 (5): 506-540.

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

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

Muralidharan K. and V. Sundararaman (2011), "Teacher Performance Pay: Experimental Evidence from India" *Journal of Political Economy* 119 (1): 39-77.

Naudet J.-D. (2006), "Les OMD et l'aide de cinquième génération" *Afrique contemporaine* 218 (2): 141-174.

Nubukpo K. (2012), "Esther Duflo, ou 'l'économie expliquée aux pauvres d'esprit'", *Blog Alternatives Economiques : L'actualité vue par Kako Kubukpo.*

Oakley A. (2000), "A Historical Perspective on the Use of Randomized Trials in Social Science Settings" *Crime & Delinquency* 46 (3): 315-329.

O'Laughlin B. (2015), "Trapped in the prison of the proximate: structural HIV/AIDS prevention in southern Africa" *Review of African Political Economy* 42 (145): 342-361.

Picciotto R. (2012), "When will the bubble burst?" *Evaluation* 18 (2): 213-229.

Pritchett L. (2012), "Impact evaluation and political economy: what does the 'conditional' in 'conditional cash transfers' accomplish?", Center for Global Development blog. online at: https://www.cgdev.org/blog/impact-evaluation-and-political-economy-what-does-%E2%80%9Cconditional%E2%80%9D-%E2%80%9Cconditional-cash-transfers%E2%80%9D.

Pritchett L., S. Samji and J. Hammer (2013) "It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space", *Harvard Kennedy School*, RWP 13-012.

Pritchett L., Sandefur J. (2015), "Learning from Experiments When Context Matters" *American Economic Review* 105(5): 471-475.

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

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

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

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

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

Scriven M. (2008), "A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research" *Journal of MultiDisciplinary Evaluation* 5(9): 11-24.

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

Schulz, K. F, I. Chalmers, R. J Hayes, and D. G Altman (1995), "Empirical Evidence of Bias" *JAMA: The Journal of the American Medical Association* 273(5): 408-412.

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

Shah N. B., P. Wang, A. Fraker and D. Gastfriend (2015), "Evaluations with impact: decision-focused impact evaluation as a practical policy making tool", *3IE*, Working Paper 25.

Sharma M. and M. Zeller (1997), "Repayment performance in group-based credit programs in Bangladesh: An empirical analysis", World Development, 25(10), pp. 1731-1742.

Teele D. L. (2014), *Field experiments and their critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences*, New Haven & London: Yale University Press.

Vivalt E. (2016), *How Much Can We Generalize from Impact Evaluations?,* Standford: Stanford University.

WHO (2014), *Preventing diarrhoea through better water, sanitation and hygiene: exposures and impacts in low- and middle-income countries*, Geneva: World Health Organization.

Woolcock M. (2013), "Using case studies to explore the external validity of 'complex' development interventions" *Evaluation* 19(3): 229-248.

Wolf, J., A. Prüss Ustün, O. Cumming, J. Bartram, S. Bonjour, S. Cairncross, T. Clasen, J. M Colford, V. Curtis and J. France (2014), "Systematic review: Assessing the impact of drinking water and sanitation on diarrhoeal disease in low and middle income settings: systematic review and meta regression" *Tropical Medicine & International Health* 19: 928-942.

Young A. (2015), "Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results", *LSE*, Working Paper.

List of recent AFD Research Papers

AFD Research Papers are available through AFD's website at www.librairie.afd.fr/researchpapers

2017-33   DAVID, A. and M.-A. MAROUANI (2017), "Migration patterns and labor market outcomes in Tunisia", AFD Research Paper Series, n° 2017-33, February.

2017-34   D'AIGLEPIERRE, R. and A. BAUER (2017), "The choice of Arabo-Islamic education in sub-Saharan Africa: findings from a comparative study", AFD Research Paper Series, n° 2017-34, March.

2017-35   AVITABILEY, C., BOBBA, M. and M. PARIGUANAX (2017), "High School Track Choice and Liquidity Constraints: Evidence from Urban Mexico", AFD Research Paper Series, n° 2017-35, March.

2017-36   BOBBA, M. and V. FRISANCHO (2017), "Learning about Oneself: The Effects of Performance Feedback on School Choice", AFD Research Paper Series, n° 2017-36, March.

2017-37   WAGNER, L. and R. D'AIGLEPIERRE (2017), "Macroeconomic Crisis, Primary Education and Aid Effectiveness", AFD Research Paper Series, n° 2017-37, February.

2017-38   BAUER, A. (2017), "Redistributive Programs' Implementation: Do Political Incentives Matter?", AFD Research Paper Series, n° 2017-38, March.

2017-39   D'AIGLEPIERRE, R. and A. BAUER (2017), "Explaining the Development of Private Education: the Effect of Public Expenditure on Education", *AFD Research Paper Series*, n° 2017-39, March.

2017-40   NOCUS, I., GUIMARD, P. and A. FLORIN (2017), "Evaluation of the « *Ecole et langues nationales en Afrique* » program: Methodological Aspects and Interim Assessment", *AFD Research Paper Series*, n° 2017-40, February.

2017-41   ALTET, M., N. SALL, M. SECK and I. OUEDRAOGO (2017), "Teacher training based on the results of research on their actual practices in the context of Sub-Saharan Africa: The OPERA project in Burkina Faso, issues and challenges", *AFD Research Paper Series*, n° 2017-41, February.

2017-42   GIRAUD, G. and M. GRASSELI (2017), "The macrodynamics of household debt, growth, and inequality", *AFD Research Paper Series*, n° 2017-42, April.

2016-43   FEUBI PAMEN, E.-P. and M. KUEPIE (2017), "An Application of the Alkire-Foster's Multidimensional Poverty Index to Data from Madagascar: Taking Into Account the Dimensions of Employment and Gender Inequality", *AFD Research Paper Series*, No. 2017-43, April.