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

Florent Bédécarrats, AFD, Isabelle Guérin, IRD-Cessma, François Roubaud, IRD-Dial

This paper is a pre-print version of Bédécarrats F., Guérin I., Roubaud F. (2019), "All That Glitters Is Not Gold. The Political Economy of Randomized Evaluations in Development", *Development and Change*, 50(3), pp.735-762.

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.

Keywords: Impact evaluation; Randomised control trials; Experimental method; Political economy; Development.
JEL codes: A11, B41, C18, C93, D72, O10

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). More broadly, Jamison (2017) considered RCTs in development as the fourth wave of randomization in empirical social science, with the introduction of the randomized assignment notion going back to the 19th century. The methods have since been tailored to poor countries' issues and circumstances. RCTs have been a resounding success, as seen from their proliferation, and the world has been quick to sing the praises of their promoters. Leading economic journals have welcomed RCTs with open arms and their reputation now transcends the disciplinary field of economics (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 using a political economy approach. Although RCTs have already been heavily criticized, the political economy of RCT implementation is still underexplored. 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. Exploring the political economy of RCTs makes for a better understanding of their limitations as a method and their success in academia, the media and among policymakers.

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. It 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 and level), 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. Many economists explain that, in the early 2000s, RCTs were regarded as a fabulous remedy by their professional community exhausted by decades of incessant controversies over potential biases in econometric identification methods (Ogden, 2017). Last and more classically, RCTs 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). 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, 2015), 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) under his belt! Yet these performances aside, are the researchers really plugged into the grassroots level (we will come back to this)?

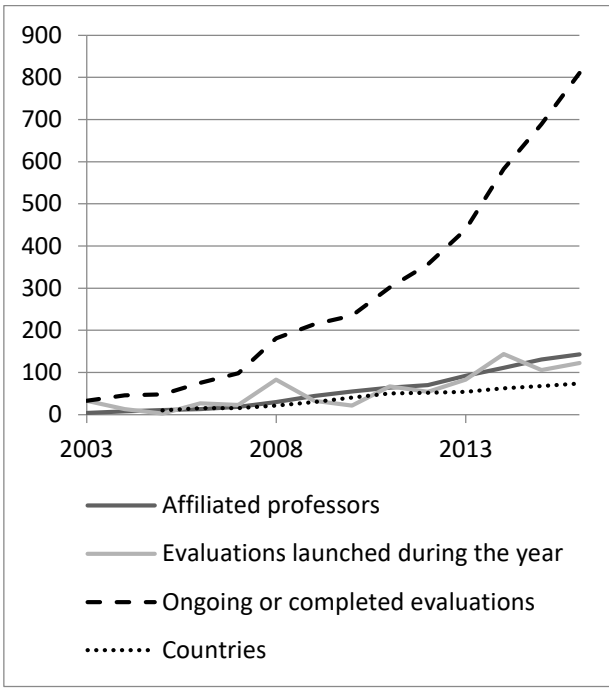
Although the World Bank appears to have less of a hard-and-fast policy when it comes to impact evaluation methodology, with an array of (nonetheless always quantitative) methods, RCTs represent nearly two-thirds of the methods used, accounting for 64% of the 368 evaluations undertaken (as at 2010). As 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 five-year 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 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).

[1] Information collected from J-PAL's website on 24 January 2017.

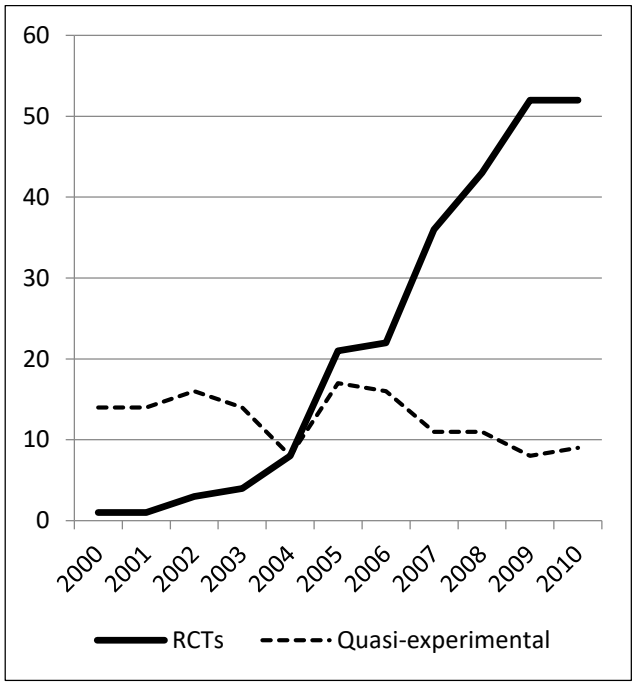
[2] In the latest version of the evaluation database we looked up on the J-PAL website, one evaluation can cover more than one sector.

Figure 1: Growth in J-PAL and World Bank impact evaluations

J-PAL (2003-2016)



World Bank (2000-2010)



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. Figure 2 presents a flattening RCT growth curve even though volumes remain high, while other indicators show no such stagnation (see Figure 1). This raises the 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

[3] International Initiative for Impact Evaluation.

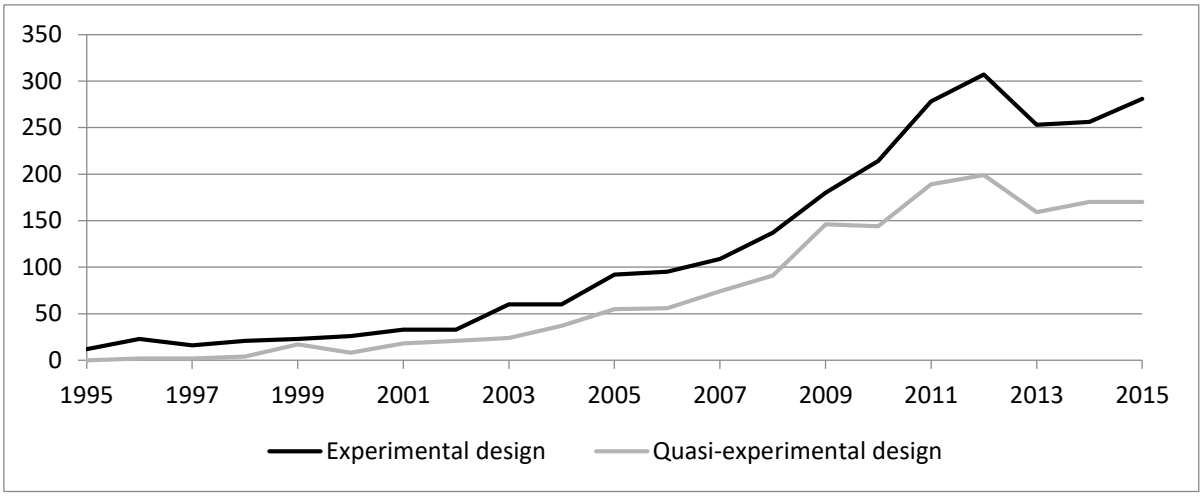
[4] www.poverty-action.org/about/annual-reports-finances, consulted on 24 January 2017.

[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*.

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.

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 before looking into the implications of their use on the ground. We show that the political economy of RCTs has first-order consequences in practice, which call into question the method's asserted theoretical properties.

IIA- 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 have a number of practical and theoretical limitations. The main theoretical criticism concerns the tension between bias and precision, where RCTs fail to make an optimal trade-off (Deaton & Cartwright, 2016; Heckman, 1991; Deaton, 2010; Ravallion, 2009; Barrett & Carter, 2010; Rodrik, 2008. 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, which is commonplace in both medicine and development. 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. Yet improving estimates entails making 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 (2016). 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. This leads Deaton and Cartwright (2016) to 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."

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. The first issue is the absence of any real random sampling for ethical or practical motives (to avoid disturbing project implementation) (Scriven, 2008; Labrousse, 2010; Deaton, 2010). Secondly, treatment variables are often too poorly specified to serve as structural estimators (Heckman & Vytlacil, 2005). Then there is the absence of blinding, which makes for various incentives for take-up … and non-take-up (Heckman *et al.*, 1998; Rosholm & Skipper, 2009). And when take-up is too low, researchers take a number of measures to "force" it, thereby turning the evaluated programme into something that is then far from a "normal" programme (Bernard *et al.*, 2012; Quentin & Guérin, 2013; Morvant-Roux *et al.*, 2014). Spillover effects and attrition effects represent yet another unsolved issue despite the many efforts to do so (Eble *et al.*, 2014).[7] These problems, long recognised in medicine, have since been evidenced by many concrete examples without RCT promoters showing any more reflexivity in their research protocols (Shaffer, 2011). That said, not all RCTs are equally undermined by such implementation distortions: the methodological properties of the RCTs conducted in development are systematically inferior to those conducted in medicine (in North and South alike) and those conducted to evaluate social policies in developed countries (Eble *et al.*, 2014).

The question of external validity is the most discussed in the literature. The focus on an "average" impact and problems capturing the heterogeneity of impacts and their distribution form a major obstacle to the relevance of the results (Ravallion, 2009; DFID, 2012; Vivalt, 2017a). 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). 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

[7] 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).

able to measure and test some intervention impacts and aspects, but they cannot analyse either their *mechanisms* or their underlying *processes*. Notwithstanding the method's limitations, the absence of theory prevents any form of understanding of the processes of change. 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).

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. O'Laughlin (2015) comes to a similar conclusion with respect to HIV in southern Africa.

RCT meta-evaluations are presented by RCT promoters as the most efficient way to overcome the external validity problem and build supposedly sound policy recommendations. Donors and evaluation groups[8] are increasingly taking up meta-evaluations and some are even dedicated users, as in the case of the Campbell collaboration. Meta-evaluations are applied to separate, but ostensibly similar experiments. They involve pooling data, estimators and associated standard errors in order to calculate an average result. This means that the basic premise of RCTs remains unchanged: interventions are supposed to produce a single effect that is entirely independent of context and circumstance. Consequently, seeing as the heterogeneity of outcomes, interventions, contexts and causality chains continues to be ignored, the basic problem of external validity remains intact (Deaton & Cartwright, 2016). In the most comprehensive meta-analysis of meta-analysis, Vivalt (2017a&b) shows that the heterogeneity of treatment effects is huge, higher than in other fields (medicine, etc.) and not rigorously addressed in the literature. Until such time as the issue of generalisability to other settings is tackled, RCTs will do little to inform decision-makers.

So 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. In the terms of reference for a study commissioned on the subject, a group of DFID managers estimated that less than

[8] See, for instance, the cases of DFID and 3ie, which produce dozen of meta-evaluations every year.

5% of development interventions are suitable for RCTs (DFID, 2012). In their more formalised paper, Sandefur and Pritchett (2013) come to a similar conclusion.[9]

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.

While some of the most prominent promoters of RCTs, including Esther Duflo, 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 – such as 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 RCT advocates' claim to offer a basket of global policies based on necessarily local RCTs. We therefore believe scale-ups 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 (Bédécarrats *et al.*, 2015). To take the RCT movement's own basic principle, the most "rigorous" way to know if RCTs have made (or will make) a difference to development outcomes would be to conduct "an RCT on RCTs" by applying RCTs to a random sample of countries/regions and comparing the results with a randomly selected control group. Such an approach is obviously not feasible, but is to be viewed rather as a thought experiment. What is puzzling at this stage is that none of the successful developing countries in recent decades (such as the Asian dragons and, more recently, China and Vietnam) have grounded their policies in RCTs or even in international academia's modern economics. Rather, they have all adopted a thoroughly pragmatic approach (see, for the case of Vietnam, Rama, 2008; Cling et al., 2013). 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.

IIB- 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 process is needed to put

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

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.

Few empirical elements are available at this level 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.[10] 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 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. 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, mentioning the evaluation's weak points would most 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 the late 1990s and the impending change in power. It was in the interest of all the players involved in the programme (researchers, promoters and decision-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

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

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. 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. 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. 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 potential. This mixed bag of priorities resulted in many compromises as much in terms of research aims, sampling and questionnaire design as in the interpretation and dissemination of the findings.

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). Following the ethical and practical demands made by the two NGOs, 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.

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. 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 health, but 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. The evaluation reports and their public presentation 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 a question that was key for the NGOs: is voluntary, paid insurance preferable to compulsory or free insurance? This question of take-up and the price elasticity of take-up was mentioned in the final academic paper, but only in passing.

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, as opposed to 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 scaled up by the government. This 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. The proliferation of RCTs in different fields can in no way make results more robust if the decisive factor for programme impact is the nature of the operator implementing it.

The article by Duflo and her co-authors (2015) was published three years after this controversy. It only mentions Bold *et al.*'s work (which has still not been published) in passing without drawing any particular conclusions from it as to the internal validity of RCTs.

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. 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 reanalysed the initial study's micro-data and found numerous discrepancies in the handling of the data (Aiken *et al.*, 2015). Once these errors were dealt with, the total effect on the reduction in school absenteeism was half that initially stated. The direct effects remained significant, but the indirect effects – the study's signature finding – no longer held.

In a second paper, Davey and his co-authors (2015) re-analysed the data using the most appropriate epidemiological statistical impact estimation methods for this type of protocol. 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 health (weight for age and height for age), which would logically be key to deworming having an effect on absenteeism. The epidemiologists more generally criticised the failure of Miguel and Kremmer's study to subscribe to the scientific standards required for public health RCTs.

The controversy fuelled by these two articles (Humphrey, 2015) unsettled those who claimed paternity of evidence-based medicine for economic randomisation. More importantly, it raised the question as to what factors were behind the scale-up of a trial with such fragile internal and external validity. 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.

IIIA- A new scientific business model

With the end of the Cold War and advent of 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 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 (Teele, 2014; Kosters *et al.*, 2015). 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 economics (Banerjee & Duflo, 2011; Karlan & Appel, 2012), 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. Their power base is grounded in a dense network of interconnected relationships, no doubt like many other leading schools of thought in economics and elsewhere. Yet more specifically, and this is what makes them unique, they have generated an entirely new scientific business model, which has in turn driven the emergence of a truly global industry. RCT researchers are young and from the inner sanctum of the top universities (mostly American).[11] They have found the formula for the magic quadrilateral by combining the mutually 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), RCT advocates have devised the means to attract young, motivated and highly skilled people. 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[12] 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 RCT advocates' 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. *Foreign Policy* and *Time Magazine* have named her among the world's most influential people and, in 2012, she was appointed advisor to President Obama on global development policy.

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 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

[11] 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.).

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

and Strategic Impact Evaluation Fund), but also bilateral (Spanish and UK cooperation agencies) and from major foundations (Rockefeller, Citi, Gates, MacArthur and Hewlett) –RCT advocates 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, co-headed by Esther Duflo, 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) is J-PAL's nerve centre. 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, 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 (Bédécarrats *et al.*, 2015: 17). 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 board of editors (Banerjee and Karlan). Esther Duflo is both the journal's editor (and 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 RCT promoters 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).

IIIB- 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.[13] 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. The ploy is particularly plain to see in that published papers based on non-experimental methods are virtually never cited (Labrousse, 2010; Nubukpo, 2012).

Despite the gathering of clouds for an outright scientific controversy (Knorr-Cetina, 1982) over RCTs, the power imbalance between stakeholders has so far prevented any full-blown storm. The debate may well be underway, but open warfare has not yet been declared. Nevertheless, although still in the minority, criticism is growing and is slowly redrawing the lines. Angus Deaton's critical voice (Deaton, 2010; Deaton & Cartwright, 2016) carries particular weight, especially given that he was awarded the Nobel Prize in Economics in 2015. This criticism is now more frequently acknowledged by RCT movement members despite its being minimised, reframed and rephrased in a way that knocks most of the stuffing out of it and points to solutions that will not undermine the method claimed as superior (Ogden, 2017).

Alongside the abovementioned deworming case, the special issue on microcredit can also be given as a typical example. It qualifies some points in response to certain criticism regarding internal and external validity. 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 (Bédécarrats *et al.*, 2015). 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, with their effectiveness challenged by meta-evaluations (WHO, 2014: 6). 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 have a limited *direct* effect (see, for example, Moffitt, 2004). The second argument 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

[13] See, on this subject, A. Deaton's interview with A. Jatteau (2016, p. 60).

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 requirements of academic publication. Arthur Jatteau (2016) has documented this constraint well in a thorough analysis of J-PAL's production and numerous interviews with RCT promoters. 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. Publication bias is probably also an issue, as suggested by Vivalt (2017a): reporting solely on larger effects (more often obtained by smaller studies) and "original" topics (with the more standard subjects left behind in the corresponding working papers). 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.

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. The RCT movement has managed to label the method as the only technique able to rigorously identify causal impacts, when RCTs actually only provide evidence of effectiveness in limited settings 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 RCT promoters' 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 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. RCT advocates 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. Credible alternatives would also be needed for the bubble to burst. This means actively pursuing thinking and action on methodological innovation.

References

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

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

Aiken A. M., Davey C., Hargreaves J.R., Hayes R.J. (2015), “Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a pure replication", *International Journal of Epidemiology*, 44(5), pp. 1572-1580.

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

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

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

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

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

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

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

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

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

Bédécarrats F., Guérin I., Roubaud F. (2015), "The gold standard for randomized evaluations: from discussion of method to political economy", *DIAL Working Paper* No. 2015-01, January.

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

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

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

Cling J.-P., Razafindrakoto M., Roubaud F. (2013), “Is the World Bank compatible with the "*Socialist-oriented market economy*"? A political economy approach applied to the case of Vietnam”, in Alary P. & Lafaye de Micheaux E. (Eds), *Political Economy of Contemporary Asia, Revue de la Régulation. Capitalisme, Institutions, Pouvoirs*, Special Issue, 13 | 1st Semester/Spring.

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

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), pp. 1581-1592.

Deaton A. (2010), "Instruments, Randomization and Learning about Development", *Journal of Economic Literature*, 48(2), pp. 424-455.

Deaton A., Cartwright N. (2016), "Understanding and Misunderstanding Randomized Controlled Trials", *NBER* Working Paper 22595, September.

Dezalay Y., Garth B.G. (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, April.

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., June, online at: https://olc.worldbank.org/content/state-economics-influence-randomized-controlled-trials-development-economics-research-and.

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

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

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

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

Eble A., Boone P., Elbourne D. (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), pp. 230-243.

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

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

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

Heckman J. J. (1991), "Randomization and Social Policy Evaluation", *NBER Technical Working Paper No. 107.*

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

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

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

IEG (2012), *World Bank Group Impact Evaluation. Relevance and Effectiveness*, World Bank, June.

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, pp. 101-130.

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

Jamison J.C. (2017), "The Entry of Randomized Assignment into Social Sciences", *Policy Research Working Paper* 8062, World Bank, May.

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, December, 542p.

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

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, October [http://www.poverty-action.org/sites/default/files/Dean%20Karlan_Testimony.pdf]

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

Kosters M., Van der Heijden J. (2015), "From mechanism to virtue: Evaluating Nudge theory", *Evaluation*, 21(3), pp. 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), pp.2- 32.

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

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

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

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

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

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

Miranda J., Sabet S., Brown A. N. (2016), "Is impact evaluation still on the rise?", *Evidence Matters: Improving development and practice*", August; 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), pp. 506-540.

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

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

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

Nubukpo K. (2012), "Esther Duflo, ou 'l'économie expliquée aux pauvres d'esprit' », Blog « L'actualité vue par Kako Kubukpo », *Alternatives Economiques*. (consulted 11 March 2013).

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

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

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), pp. 342-361.

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

Pritchett L. (2012), "Impact evaluation and political economy: what does the 'conditional' in 'conditional cash transfers' accomplish?", Center for Global Development blog. Available from 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 [Accessed 25 March 2017].

Pritchett L., Samji S., Hammer J. (2013) It's All About MeE: Using Structured Experiential Learning ("e") to Crawl the Design Space, Harvard Kennedy School, RWP13-012.

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

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

Quentin A., Guérin I. (2013), « La randomisation à l'épreuve du terrain. L'expérience de la micro-assurance au Cambodge", *Revue Tiers Monde*, 1(213), pp. 179-200.

Rama M. (2008), "Making Difficult Choices: Vietnam in Transition", *Commission on Growth and Development*, Working Paper No. 40, Washington D.C., The World Bank.

Ravallion M. (2009), "Evaluation in the practice of development", *The World Bank Research Observer*, 24(1), pp. 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., Skipper L. (2009), "Is labour market training a curse for the unemployed? Evidence from a social experiment", *Journal of Applied Economics*, 24, pp. 338–365.

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

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

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), pp. 408-412.

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

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

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. (2017a), "How Much Can We Generalize from Impact Evaluations?", Stanford University, 23 September.

Vivalt E. (2017b), "How Much Can Impact Evaluations Inform Policy Decisions?", Stanford University, 30 July.

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), pp. 229-248.

Young A. (2016), "Channelling *Fisher*: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results", LSE Working Paper, February.