UMR 225 IRD - Paris-Dauphine

WORKING PAPER WP/2015-01

The gold standard for randomised evaluations: from discussion of method to political economy

Florent BEDECARRATS
Isabelle GUERIN
François ROUBAUD

UMR DIAL 225
Place du Maréchal de Lattre de Tassigny 75775 • Paris Cedex 16 •Tél. (33) 01 44 05 45 42 • Fax (33) 01 44 05 45 45
• 4, rue d'Enghien • 75010 Paris • Tél. (33) 01 53 24 14 50 • Fax (33) 01 53 24 14 51
E-mail : dial@dial.prd.fr • Site : www.dial.ird.fr

THE GOLD STANDARD FOR RANDOMISED EVALUATIONS: FROM DISCUSSION OF METHOD TO POLITICAL ECONOMY

Florent Bédécarrats
AFD
75012 Paris, France
bedecarratsf@afd.fr

Isabelle Guérin
IRD-CESSMA
75013 Paris, France
isabelle.guerin@ird.fr

François Roubaud
IRD, UMR DIAL, 75010 Paris
PSL, Université Paris-Dauphine,
LEDa, UMR DIAL, 75016 Paris, France
roubaud@dial.prd.fr

Working Paper UMR DIAL
February 2015

Abstract

This last decade has seen the emergence of a new field of research in development economics: randomised control trials. This paper explores the contrast between the (many) limitations and (very narrow) real scope of these methods and their success in sheer number and media coverage. Our analysis suggests that the paradox is due to a particular economic and political mix driven by the innovative strategies used by this new school's researchers and by specific interests and preferences in the academic world and the donor community.

Key words: Impact evaluation, Randomized control trial, Experimental method, Methodology, Political economy, Development.

JEL Code : A11, B41, C18, C93, D72, O10

Résumé

La dernière décennie a vu l'émergence d'un nouveau champ de recherche en économie du développement : les méthodes expérimentales d'évaluation d'impacts par assignation aléatoire. Cet article explore le contraste entre d'une part les limites (nombreuses) et la circonscription (très étroite) du champ réel d'application de ces méthodes et d'autre part leur succès, attesté à la fois par leur nombre et leur forte médiatisation. L'analyse suggère que ce contraste est le fruit d'une conjonction économique et politique particulière, émanant de stratégies novatrices de la part des chercheurs de cette nouvelle école, et d'intérêts et de préférences spécifiques provenant à la fois du monde académique et de la communauté des donateurs.

Mots Clés : Evaluation d'impact, méthode expérimentale, Essai randomisé, Méthodologie, Economie politique, Développement.

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 and already used for public policy evaluation in developed countries (mainly the United States; Oakley, 2000) since the 1960s have been tailored to poor countries' issues and circumstances. They have been a resounding success, as seen from their proliferation, and the world has been quick to sing the praises of their promoters. The leading economic journals have welcomed RCTs with open arms in a surge of published articles, the most recent example of which is a special issue on microcredit (Banerjee *et al.*, 2015). 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 or Marseille Schools of Economics, etc. 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 (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, such as Esther Duflo, are looking to stretch RCTs well beyond their methodological scope in a bid to establish a list of all good and bad development policies. The grounds for this upscaling ambition would appear to be 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 and their theoretical advantages, especially compared with other existing evaluation methods. 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, in a move to establish the extent of their validity in the more general development arena. The third and last section takes a political economics angle to understand how the different players interact, the power games and struggles at work, and who gains from them. It seeks to explain why these methods enjoy political and academic credibility far beyond their real scientific reach. 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. Lastly 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).

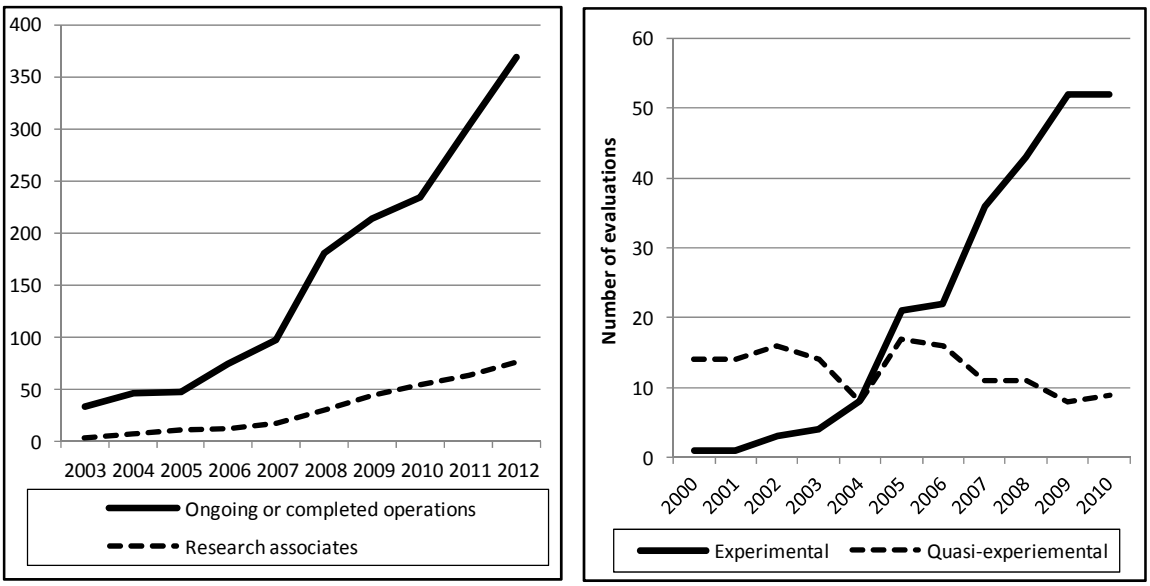
In addition to these generic (and theoretical) plus points, there are other reasons to commend the replication and proliferation of RCTs in development. We note the three main reasons. Firstly, RCTs have put their finger on a blind spot in both national policies and policies driven by official development 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, 2006; 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; 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 Lateef 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 specialises solely in RCTs and is recognised as a quality label in the field. It organises training courses and exchanges of practices with a network of 100 affiliated professors[1] and a host of researchers. 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 March 2014, ten years after its establishment, J-PAL was posting a total of no less than 573 evaluations (ongoing or completed) in 61 countries, with exponential growth over the years. Africa is its leading ground (with 176 trials), way ahead of South Asia (123, mainly in India) and Latin America (108). Top focal points are microfinance (177) followed by the social sectors (health and education, with 139 and 129 respectively) and governance, which is a growing focus with 118 trials. Esther Duflo plays a key role in the Lab, having worked on no less than 46 trials (14 of which are ongoing), although she is outdone by Dean Karlan who has 78 trials (36 ongoing) under his belt! Yet these performances aside, are the researchers really plugged into the grassroots level (we will come back to this)?

Growth in J-PAL and World Bank impact evaluations

J-PAL (2003-2012) **World Bank (2000-2010)**



Sources: J-PAL website, February 2013; IEG, 2012.

[1] Information available on J-PAL's website (http://www.povertyactionlab.org) consulted on 26 November 2014.

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 become king of the castle, they are crowding out the other approaches, as shown by the figure below. 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.

Development impact evaluations have been a dazzling success story in recent years, to the point of becoming quite the industry (over 1,000 ongoing or completed evaluations reported by White, 2014). Local and international initiatives have followed hot on their heels (such as 3ie and NONIE[2]). Experimental methods have gradually swallowed up the other methods and are now seen as the "gold standard" by donors and researchers alike (Pritchett & Sandefur, 2013). They have spread to the four corners of the globe, using considerable resources (hundreds of millions of dollars) and generating dozens of academic papers and a myriad of best practice manuals. Yet why is this striking phenomenon (called a pro-impact evaluation movement) all the rage?

Is this systematic use of RCTs scientifically sound and politically expedient? Two questions pose problems. First, there is the alleged intrinsic superiority of RCTs over any other method (Duflo *et al.*, 2007). Second, there is 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).

II. Implementation of the method: small scope

RCTs have three kinds of limitations: i) biases that can undermine the reliability of the results obtained (internal validity); ii) the limited extent to which findings can explain situations other than the given case study (external validity); and iii) holes in the RCT scale-up plan. We address them in order.

Internal validity

Advocates of the use of RCTs in development economics imported the method from the medical world without a thought for 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 chose to overlook the controversies that had raged through decades of development economics debates.

RCTs are no exception to the habitual "tweaking" of research protocols, especially in social science. In many cases, practical constraints and ethical considerations carry us far from the ideal world of laboratory trials. These divergences from the theory can challenge the very principle of random sampling and absence of selection bias and hence the superiority of RCTs over other statistical and

[2] 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 the end of 2013, it had funded 150 impact evaluations and systematic reviews for a total of 66.6 million dollars. NONIE (Network of Networks for Impact Evaluation) takes in all the networks promoting impact evaluations.

econometric techniques (Scriven, 2008; Labrousse, 2010; Deaton, 2010). Field partners sometimes refuse to randomly sample beneficiaries for ethical reasons, thereby forcing research teams to opt for other selection methods such as alphabetical sampling. Yet this can generate bias since many resources are also allocated alphabetically (Deaton, 2010). There are also times when the managers of projects under impact evaluation insist on specific selection protocols so as not to disturb their intervention methods. In a microhealth insurance project in Cambodia, for example, participants were drawn from a lottery at a village meeting. Yet there is a good chance that the people who attend village meetings are more available, more curious and open to innovation, closer to the village leader, more socially integrated, in poorer health, etc. Funnily enough, this question is not discussed (Quentin & Guérin, 2013). In certain cases, then, a question mark hangs over the neutrality of the selection protocols and their real capacity to eliminate the differences in "unobservable characteristics" between treatment and control group that RCTs are supposed to erase.

Random sampling protocols also sometimes meet with resistance or lack of interest from beneficiaries, especially when interventions rely on people willingly coming on board or imply costs for the participants. It then becomes very hard to ensure that the randomly selected people will subscribe to it or that those who are excluded by the lottery will not have access to it by other means.[3] Yet when recipient ownership is lower than expected and incompatible with the requirements of statistical precision, research teams do not think twice about bending the participation rules. In two recent studies, one on the abovementioned microinsurance projects and the other on microcredit in rural Morocco, special information campaigns, fieldworker incentives and discount rates were organised at the request of the researchers (Bernard *et al.*, 2012; Quentin & Guérin, 2013; Morvant-Roux *et al.* 2014). In these situations, the evaluated programmes are then far from "normal" programmes.

The principle of randomisation assumes that the control group (a population and/or a territory) remains free from any intervention throughout the study. Yet in thriving or competitive sectors (such as microcredit), it is unrealistic to expect control groups to remain untouched by any intervention and impervious to its effects, as the provisions of the research protocol would have it. Attrition is also problematic. Participants frequently drop out of the programme (the treatment) over time. Another inevitable phenomenon is the loss of sampled individuals from one survey wave to the next for various reasons (decease, migration, refusal to respond, etc.), which are not necessarily the same across treatment and control groups. Yet where there is more attrition among certain types of participants/respondents, this introduces selection bias after the fact. Even when researchers try to take account of attrition effects in their analyses, which is far from systematic (Elbe *et al.*, 2014), it is very hard to correct the resulting bias.[4] Aside from in certain highly specific cases, it remains hugely difficult to account for spillover effects and externalities of all kinds, such as attrition, despite attempts by RCT experts to improve their definitions of the different populations for whom the impacts are estimated (ATE: Average Treatment Effect; ATT: Average Treatment on the Treated; LATE: Local Average Treatment Effect, etc.; Duflo *et al.*, 2007).

[3] See, for example, Heckman *et al.* (1998); Rosholm & Skipper (2009).

[4] 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 (Elbe *et al.*, 2014).

Moreover, to prevent distortions induced by local arrangements (e.g. where a beneficiary drawn at random offers or sells his or her participation to a neighbour), random sampling also calls for a great deal of safeguards, very close cooperation with partners on the ground, and suitable survey interviewer training and motivation. More broadly speaking, research protocols are so hefty (complex lotteries with multi-round surveys, samples thousands strong, and homogenisation of intervention methods), time-consuming[5] and expensive that associations of players need to be formed with both the evaluated project operators and the local research teams in charge of conducting the surveys, and sometimes with other correspondents imposed or suggested by the donors (e.g. governments). A study on the actual implementation of an evaluation for a microhealth insurance programme in Cambodia shows the complexity of these player alliances and especially the compromises and many adjustments that have to be made with respect to sampling, research aims, questionnaire design and interpretation of the findings, all making changes to the initial protocol that cause the evaluation to depart from its sound theoretical properties (Quentin & Guérin, 2013). Although this statistical and institutional tweaking is commonplace, a standard sleight of hand among randomistas,[6] it is generally swept under the carpet when the findings are published even though it has decisive repercussions on the nature of the protocols and hence the internal validity of the studies. The most emblematic case in point is the evaluation of the Progresa programme by the International Food Policy Research Institute (IFPRI). 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 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 all RCTs. Yet this did not prevent the subsequent publications from presenting the sampling protocol as truly experimental. Only by gradually papering over 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 for evaluating the impact of development programmes and also, in the Mexican case, keeping the programme going following the upcoming presidential elections in the late 1990s and the impending change in power.

Yet aside from the fact that often no mention is made of it, this frequent bending and stretching of the theoretical protocols used to set up RCTs would not be so problematic if it were seriously taken

[5] Even for short-term surveys (such as one year), studies can last four to five years due to the unwieldiness of the preparation protocols (and the number of studies the RCT teams are handling).

[6] See, for example, the discussions on the World Bank's impact evaluations blog (http://blogs.worldbank.org/impactevaluations/). In his post on *The impact evaluation roller coaster*, Goldstein (2011a) describes the problems associated with the following constraints: arguing your case with many people at once to convince the operational team that an RCT is vital and negotiating the protocol to secure adequate statistical power; dealing with government department replacements to keep the project manager on board; adjusting the protocol when implementation is ahead of schedule; sorting things out when the wrong list is used for the draw in public, etc. In another post, he talks about the need to work with local community insiders to facilitate community entry (Goldstein, 2014), but without asking questions about any bias this may induce. He also wonders about "counter-placebo" effects in the control group (Goldstein, 2011b). In the same vein, Ozler (20123 discusses the problems associated with random draws and response biases in questionnaires (which are nonetheless valid for any questionnaire-based survey).

into account, which is far from the case. For example, Eble *et al.* (2014) conduct a meta-analysis comparing all the RCTs published in the top 50 international economics journals from 2001 to 2011 (54 articles) with a same-size sample representative of medical RCTs published in the "top" three medical journals, taking their methodological characteristics as the benchmark standard. They show that the first set of RCTs perform significantly less well than the second set across all criteria considered. The economic RCTs are systematically more exposed to the risk of each of the six sources of bias identified in the medical literature.[7] Even though the procedures to control for performance and detection biases (introduction of placebos and blinding) are often harder to tackle in social programmes, an effort still needs to be made to discuss their importance. All in all, therefore, economic RCTs are more subject to overestimating the effects of the studied treatment. This raises questions as to the robustness of the results obtained. Economic RCTs are also less rigorous when it comes to documenting these potential biases and ultimately seeking to correct them. Moreover, the quality of the more recent RCTs, such as those published in the top three journals (versus the 47 others), is not any better.

These shortcomings cannot be explained by the fact that the risks of bringing the treatments to scale would be higher in the medical sphere than in economics. Both cases entail potentially lethal effects, especially in developing countries. In addition, the quality of the economic RCTs is lower than in developed countries, which is not the case in the medical field (Eble *et al.*, 2014). These weaknesses have more to do with a more casual approach due to less of a need for scientific quality and laxer peer review procedures.

External validity

Turning to external validity, the focus of RCTs is generally so narrow that there is no telling whether anything can really be learned from them beyond the immediate intervention studied. Statistical precision dictates that RCTs need very large population samples even to answer the basic question as to whether the intervention has improved the beneficiaries' situation. So they are not well placed to address the heterogeneity of the effects, which most pass over to concentrate on average impacts (Heckman, 1991). Yet the effects of development policies can differ enormously depending on the different characteristics of their beneficiaries and how the policies are implemented locally.[8] Does the project work better in certain settings than in others? This question is key to service improvement and project roll-out. The problem here also concerns the fallacy that an average impact can adequately capture the complexity of all the effects (Ravallion, 2009; DFID, 2012). In microfinance, for example, it makes little sense to consider an average impact when there is such a wide range of microcredit service supply arrangements and such a diverse demand for this type of service depending on the user (Bouquet *et al.*, 2009; Mosley & Hulme, 1998) and the place of establishment (Morvant *et al.*, 2014). This problem is due to technical and, ultimately, financial constraints. Studying a diversity of effects with adequate statistical power implies even larger samples and hence higher costs.

[7] That is selection, performance (change of behaviour when the evaluation subjects know their treatment status; also called the Hawthorne effect and the John Henry effect), detection (change of behaviour when the evaluators know the subject's status), attrition, reporting and sample size biases (see Elbe *et al.*, 2014).
[8] See, for example, Morvant-Roux *et al.* (2014), in the case of microcredit.

RCTs are often very small in terms of space – targeting specific geographic areas – and time so as to keep trial parameters under control. Cost and attrition are such that RCTs rarely study the impact of an intervention for more than two years. Some are even limited to one year, especially where take-up is too low and beneficiary dropout rates are high. Randomistas are not unaware of this constraint[9] and some evaluations manage to cover longer periods of time, but this remains the exception to the rule.[10] This time constraint often leads researchers to focus on mid-term results rather than long-term outcomes. For example, the evaluation of a farmer poverty reduction programme measured the increase in the use of fertilizer rather than farmers' earnings or level of poverty (Labrousse, 2010; Eble *et al.* 2014). The evaluation of a microhealth insurance programme measured the reduction in the families' level of debt rather than their state of health (Quentin & Guérin, 2013). The question could also be put as to whether the impacts stand the test of time (Labrousse, 2010). Impacts observed in the short run can change, if not completely reverse in the medium to long run, for all sorts of reasons. In microcredit, for example, the negative effects of over-indebtedness observed in several world regions in recent years (Guérin *et al.* 2013) often appear after a number of microloan cycles. More generally speaking, rarely are development project trajectories linear and this largely restricts the validity of the conclusions obtained after a short lapse of time (Woolcock, 2009).

Many development programmes have knock-on effects, composition effects and general equilibrium effects, i.e. they have positive or negative effects on a region or a sector that extend far beyond the beneficiaries of the programme itself. Such is the case, for example, with job programmes. They can raise total earned income (Ravallion, 2009), but they can also trigger saturation and substitution effects. A typical example of this is found in microcredit where the businesses created by these loans can cause saturation on local markets or reduce demand for competing firms, if not drive them to bankruptcy. This type of effect is highly frequent (Bateman, 2010), but is very often neglected by randomised studies.

Upscaling an intervention designed for a town or village to a region or even a country is by no means a straightforward matter, and nationwide roll-outs of small-scale local programmes are consequently problematic. Scaling up implies more than just technical considerations (externalities, spillover effects, saturation, general equilibrium effect, etc.). It is also a question of political economics. For example, Bold *et al.* (2013) show that where a contract teacher programme in Kenya was found to have a positive impact on pupils' levels of education when applied by an NGO on a small scale (RCT), this positive effect disappeared when the programme was scaled up nationally and implemented by the government. This study, based on the use of a highly original RCT (the first of its kind to test organisational and political economy effects) taking advantage of the programme's national rollout, concludes that the absence of effects following upscaling was due to the change of project operator: from carefully selected, highly motivated NGOs to the bureaucratic structures of government and associated organisations (unions). This bias could even be systematic in the event of a correlation between the places/people/organisations agreeing to implement RCTs and the estimated impacts (Pritchett & Sandefur, 2013). We believe that Acemoglu (2010) has a particularly decisive argument to make when he discusses political economy responses to large-scale programmes from groups who see their rents threatened by reforms. This is a key question when seeking to go to scale with the results of RCTs on locally conducted policies.

[9] See, for example, Goldstein's post (2011c).
[10] See, for example, Baird *et al.* (2012) on deworming.

Looking at their scientific reach, RCTs might be able to measure and test some intervention impacts and aspects, but they cannot analyse either the *reasons* for them or their underlying *processes* (Ravallion, 2009). The mechanisms that drive a particular intervention to produce an outcome remain a blind spot for RCTs (Rao & Woolcock 2003; Hulme 2007), due in part to their not being placed in context (Pritchett & Sandefur, 2013). Understanding (rather than just measuring) the impacts of an intervention would call for an analysis of the phenomena studied from every angle and an examination of the complexity of the causal links and the many, dynamic and contradictory interactions between the different entities addressed in a contextual, location-specific way. It would also require analyses at both meso level (covering context, the institutional setting of the actions taken, etc.) and micro level (comprehensive analysis of household behaviour) and studies of the complex links between different entities and different levels. Pritchett and Sandefur (2013) take two illustrative examples in the economics of education (class size effects and gains from schooling) to 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. On a more general level, they set out to categorically prove that the claim of external validity for RCT-estimated impacts is necessarily invalid. When context effects are taken into account, there is a trade-off to be made between the choice of a "good estimate" (internal validity of an RCT) obtained in a different place to that in which it is intended to be applied (another or larger – typically national – geographic area) and a "bad estimate" (i.e. not taking account of selection into treatment) from the right place. This consequently directly challenges the idea that RCTs are the best possible method (the gold standard), as underscored by the authors in question. The road to mistaken policy recommendations is paved with the belief that one set of results is more rigorous than those obtained by other methods. The use of RCTs ultimately calls for many conditions found only in specific cases called tunnel-type programmes (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. 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.[11]

While some of the most prominent promoters of RCTs, especially 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).[12]

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

[12] The J-PAL website features a special page on scale-ups divided into sub-headings (governance, health, etc.). The number of beneficiaries of programmes designed from pilots evaluated by J-PAL (a total of 202 million

Nevertheless, due to their abovementioned characteristics, RCTs focus on small, relatively simple and easily actionable set-ups that 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. François Bourguignon, a prominent researcher who has largely contributed to promoting RCTs, thinks this proposal is crazy and scientifically impossible.[13] Pursuing this gargantuan project is therefore at best impetuous, but is more probably driven by interests that should be identified.

III.- Political economics of a scientific enterprise

To understand the disconnects between the method's limitations and its huge credibility, both in the academic and the political field, we first have to consider the balances of power at work that go towards forging collective preferences for a given 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. This demand is twofold: it comes from both the donor community and the academic world.

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. The end of the Cold War saw the political sphere easing its grip on official development assistance (ODA). Cold War technical and financial cooperation was often merely another pawn in bloc rivalry. As the Berlin Wall fell, so did cooperation's subordination to real politik. In the new post-modernist world, ODA promoters have found themselves in the hot seat as the aid crisis, MDGs and New Public Management have summoned them to the stand to prove their utility (Naudet, 2006).

---

people in January 2014) appears to be the main indicator for the success of these scale-ups. The narrow spectrum covered also raises questions about the real reach of these scale-ups: police skills training for the "political economy and governance" sub-heading, deworming and remedial education for "education", and free insecticidal bednets for "health". See http://www.povertyactionlab.org/scale-ups (consulted on 28 January 2015).

[13] 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. The "pure, scientific evidence" on all that is concerned with development is simply completely impossible. We have to live with this imperfect knowledge." *(our underlining)*

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 and then in Busan in 2011. 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 means), elegant use of mathematical theory (guarantee of scientificity) and focus on the poor (household surveys).

The climate on the academic side, first and foremost 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 (Berndt, 2014; Labrousse 2013). Their simplicity makes them easy for policymakers to understand, lending them appeal as a vehicle for informing public decision-making.

RCT advocates point up their simplicity and ease of understanding and appeal to policymakers. The evaluation of the *Progresa* programme in Mexico (Skoufias & Parker, 2001) formed a prototype for this method and a textbook example of its performance capabilities. The positive results of this evaluation were put forward as an argument to sustain and roll out this conditional cash transfers (CCT) measure, which would probably have been axed following the election of the opposition party to the Mexican government. This model victory by scientific evidence over the vicissitudes of politics placed an effective argument in the hands of those advocating that these methods should become the basis for development policymaking,[14] an argument hotly disputed as mentioned above (Faulkner, 2014).

The World Bank was also 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, toward 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 – Development Impact Evaluation Initiative), 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

[14] The case of *Progresa* (later called *Oportunidades*) is emblematic in that it combines a policy type (CCTs) with an impact evaluation type (RCTs) held up as mutually reinforcing success stories each in their respective fields.

policymakers. In addition to its arguments, the report also acted as self-serving advocacy since it increased the profile of and demand for studies from many of its authors, first and foremost RCTs.

The wave of support for RCTs has also surged with the emergence of a new generation of researchers. They are young and from the inner sanctum of the top universities (mostly American). 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 taught in classic (face to face) and new forms (MOOC), the randomistas have devised the means to attract young, motivated and highly skilled resources.[15] In an intense whirl of communication and advocacy, backed by a plethora of press and para-academic media (policy briefs, blogs, outreach forums, happenings, etc.), they give the welcome impression of researchers stepping out from their ivory tower. Their modest, grassroots position[16] oozes commitment, empathy and impartiality.

A study of the career paths of high flyers and their networks can be a compelling angle to understand the emergence, decline and transnational spread of scientific and policy paradigms (Dezalay & Garth, 2002). It is not our intention here to embark on such an undertaking with respect to RCTs, which would form a research programme of its own. We will settle instead for an outline. The figure of Esther Duflo is the most illustrative example of this movement. 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, but 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, created and funded by AFD (French Agency for Development). 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 their studies. By managing to create their own funding windows – mainly multilateral (the World Bank initiative for the development impact evaluation, the international impact evaluation initiative, the African

[15] 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.

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

Development Bank and the Strategic Impact Evaluation Fund), but also bilateral (Spanish and UK cooperation agencies) and from major foundations (Rockefeller, Citi and Gates) – 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, co-headed by Esther Duflo, is the most emblematic and accomplished example of this new scientific business model. The J-PAL laboratory itself, set up by A. Banerjee and E. Duflo,[17] is one of the MIT Economics Department's research centres. These institutional roots, with one of the most prestigious 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. In addition to its RCT communication and advocacy role,[18] this non-profit organisation works to scale up and replicate randomised control trials once they have been tested by J-PAL. And so it is that the alliance of the two institutions is set up to pull off 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 78 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 (and 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.[19] 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-signs a paper and two of them are members of the 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 publications. 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

[17] The third founder member, S. Mullainathan, is currently a professor at Harvard.

[18] Unlike J-PAL, which needs to comply with academic decorum, IPA can state its aims in marketing terms: "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/).

[19] 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*.

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.[20]

Given these circumstances, it is easier to understand why criticism of RCTs is greeted with hostility and fiercely opposed by RCT promoters.[21] A number of strategies are employed to establish the monopoly and supremacy of RCTs. Alternative methods are discredited as RCTs assume the scientific monopoly (Harrison, 2011). Critical voices have long been kept 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 recognised 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 redundant.[22] 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. A few microfinance examples provide a good illustration of this tendency to pass off as original theories already largely explored in the literature.[23]

Group lending versus individual lending. 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 the prestigious 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; Harper & Dichter, 2007). The question put regarding 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

[20] Note that teachers in the top American universities earn incomes incommensurable with French academia. Moreover, the abovementioned World Bank research evaluation report considers that Bank officials are poorly paid compared with their academic counterparts and that this is responsible for a brain drain.

[21] Incidentally, but symptomatically, the title of the 2012 AFD-EUDN Paris conference on evaluation, *Evaluation and is Discontents* (AFD, 2012) did not go down well with certain randomistas who asked for it to be changed. It was particularly frowned upon since it put a dent in the hagiographic endeavour still in the making in France and was considered to be in especially bad taste coming from AFD, the main potential (and actual) donor for RCTs in the development field in France. It was moreover AFD that had contributed to the creation and funding of the "Knowledge Against Poverty" Chair at the Collège de France, awarded to Esther Duflo on its launch in 2009.

[22] 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 economics.

[23] 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.

conclusions (Godquin, 2004; Sharma & Zeller, 1997; Gonzalez-Vega *et al.*, 1996). Yet these 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 (2011) 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 here. 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.*, 2009). 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.[24]

Despite the gathering of storm clouds for a real scientific controversy (as defined by Callon *et al.*, 2001; see also Knorr-Cetina, 1982) on the RCT issue, the power imbalance between stakeholders has so far prevented any full-blown row. The debate may well be bubbling under, but open warfare has not yet been declared. Note, however, that criticism is growing and gradually redrawing the lines. Here again, the special issue on microcredit could be mentioned by way of an example. The issue is 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.*, 2015). 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). Yet this remains but a slight shift in position. Basically, the two main founding principles are still there: 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.*, 2015, p. 2) allegedly resolves the criticism of the method's external validity.

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 economics 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 step forward has come with two steps

[24] See IPA's description of its strategy on its webpage: http://www.poverty-action.org/about, consulted on 22 January 2015.

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.

Among the possible extensions of this paper, two lines of inquiry appear to be promising: one analytical and the other methodological. On the first front, our political economics approach is worth rounding out with historical and sociology of scientific knowledge studies. 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 & Woolgar, 1978; 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.

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.

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 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 evaluation methods.

References

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

AFD (2013), *Politique d'évaluation*, AFD, Paris, October.

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

Baird, S., Hicks, J., Kremer, M., Miguel, E. (2012),"Worms at Work: Long-run Impacts of Child Health Gains", *Unpublished* (http://scholar.harvard.edu/files/kremer/files/klps-labor_2012-03-23_clean.pdf), [consulted on 23 December 2014].

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

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

Banerjee, A., Duflo, E., Glennerster, R., Kinnan, C. (2009), "*The Miracle of Microfinance? Evidence from a Randomized Evaluation*", Working Paper J-PAL.

Banerjee, A., He, R. (2008), "Making Aid Work", in W. Easterly and N. Birdsall (eds.), *Reinventing Foreign Aid*, Cambridge (MA): MIT Press.

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

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

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

Bateman, M., (2010), *Why doesn't Microfinance Work? The Destructive Rise of Local Neoliberalism*, London: Zed Books.

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

Bédécarrats, F. (2012), « L'impact de la microfinance : un enjeu politique au prisme de ses controverses scientifiques », *Mondes en développement*, No. 158, pp. 127142.

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

Berndt, C. (2014), "Behavioral economics, experimentalism and the marketization of development", (submitted)

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.

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

Cling, J.-P., Razafindrakoto, M., Roubaud, F., (Ed.) (2003), *New International Poverty Reduction Strategies*, London: Routledge.

Deaton, A. (2009), *Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development*, Cambridge: National Bureau of Economic Research.

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

Dezalay, Y., Garth, B. (2002), *The internationalization of palace wars: lawyers, economists, and the contest to transform Latin American States*, Chicago: University of Chicago Press.

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.

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

Durand, C., Nordmann, C. (2011), « Misère de l'économie du développement », *La Revue des livres* No. 1, Sept/Oct. (http://www.revuedeslivres.fr/misere-de-leconomie-du-developpement-cedric-durand-et-charlotte-nordmann/)

Duflo, E. (2010), *Lutter contre la pauvreté*, tomes 1 et 2, Paris: Le Seuil.

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, 205–231.

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

Easterly, W. (2006), The *White Man's Burden*: Why the *West's* Efforts to Aid the Rest have done So Much Ill and So Little Good. New York: Penguin Press.

Elbe, 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): 230-243.

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.

Goldstein, M. (2014), "Notes from the field: community entry and seatbelts**",** Development Impact: News, views, methods, and insights from the world of impact evaluation, (http://blogs.worldbank.org/impactevaluations/notes-field-community-entry-and-seatbelts), [consulted on 23 December 2014].

Goldstein, M. (2011a), "The impact evaluation roller coaster", Development Impact: News, views, methods, and insights from the world of impact evaluation. (https://blogs.worldbank.org/impactevaluations/the-impact-evaluation-roller-coaster) [consulted on 24 December 2014]

Goldstein, M. (2011b), **"**Is it the program or is it participation? Randomization and placebos**",** Development Impact: News, views, methods, and insights from the world of impact evaluation, (http://blogs.worldbank.org/impactevaluations/is-it-the-program-or-is-it-participation-randomization-and-placebos), [consulted on 23 December 2014].

Goldstein, M. (2011c), **"**In the long run...**",** Development Impact: News, views, methods, and insights from the world of impact evaluation**, (**http://blogs.worldbank.org/impactevaluations/in-the-long-run), [consulted on 23 December 2014].

Guérin, I., Morvant-Roux, S., Villarreal, M., (Ed.), (2013), *Microfinance, Debt and Over-indebtedness. Juggling with Money*, London: Routledge.

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

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

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

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

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, Massachusetts: Harvard University Press.

Latour, B., Woolgar, S. (1978*), Laboratory Life, the Construction of scientific facts*, Beverly Hills/London: Sage Publication.

Moyo, D. (2009), *Dead Aid: Why aid is not working and how there is another way for Africa*, London: Allen Lane.

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: 302-312

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

Ozler, B. (2012), "When Randomization Goes Wrong...", *Development Impact: News, views, methods, and insights from the world of impact evaluation*. (http://blogs.worldbank.org/impactevaluations/when-randomization-goes-wrong) ([consulted on 23 December 2014].

Ozler, B. (2013), "Economists have experiments figured out. What's next? (Hint: It's Measurement)**",** *Development Impact: News, views, methods, and insights from the world of impact evaluation*.
(http://blogs.worldbank.org/impactevaluations/economists-have-experiments-figured-out-what-s-next-hint-it-s-measurement) [consulted on 23 December 2014].

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) : 179-200.

Ravallion, M. (2009), "Evaluation in the practice of development", *The World Bank Research Observer*, 24(1): 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): 11–24.

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

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

White, H. (2014), "Current Challenges in Impact Evaluation", *European Journal of Development Research*, 26(1): 18-30.