Learning from Randomized Controlled Experiments

The Narrative of Scientificity, Practical Complications, Historical Experience

By Agnès Labrousse

Randomized controlled experiments hold out the promise of heightened scientificity and new forms of social action. In this essay, Agnès Labrousse points out some of the practical limits of these experiments and situates them within a longer history of social experimentation and governing by evidence.

Since the early 2000s, randomized experiments have been an undeniably fashionable methodology, thanks notably to the work of J-PAL (Jameel-Poverty Action Lab) and its founders, Esther Duflo and Abhijit Banerjee. Their work, which is frequently acclaimed, has attracted the attention of the most prestigious economic journals, the mainstream press, and this website (Mayneris, 2010; L'Horty and Petit, 2011; Bérard & Valdenaire, 2013). These experiments are now promoted by international organizations such as the World Bank or the Bill and Melinda Gates Foundation. They first took off in France with the Fonds d'Expérimentation pour la Jeunesse (Youth Experimentation Fund, or FEJ), which was launched by Martin Hirsch[1] (Bureau et al., 2013). Taking as our starting point the work of Duflo's team—which proposes one of several possible approaches to these experiments—we will then broaden our perspective by discussing social experimentation's longer history.

Often presented as revolutionary, this methodology purports to be highly scientific, transparent, and efficient in terms of social action (1). Yet recent scholarship has identified the major shortcomings that become evident when such experimentation is implemented and

[1] Martin Hirsch, the former head of the French charity Emmaus, served as High Commissioner for Active Solidarity against Poverty between 2007 and 2010.

utilized (2). It is striking that these very limitations had already been brought to light in the United States in the 1970s, during the last wave of randomized experimentation. Indeed, we are dealing with a recurrent project of governing by evidence, which, ever since the first social randomized controlled experiments of the 1920s, has given way to cycles of enthusiasm and deception among scientific and political actors (3).

1. A Claim to Scientificity, Simplicity and the Renewal of Public Action

Randomized experiments promise three things: scientificity, transparency, and a renewal of social action in general and public action in particular.

A Claim to Scientificity

The introduction of a methodology tested in the medical field—randomized controlled clinical trials— into the fight against poverty and public policy evaluations has supposedly resulted in increased rigor: "anti-poverty policies are evaluated with the rigor of clinical trials" (Duflo, 2009). The European website of J-PAL, the largest global poverty alleviation laboratory, says in a similar vein: "J-PAL based its reputation on the exclusive use of controlled experiments that make it possible to produce results of exceptional scientific rigor and quality."

A Claim to Simplicity and Legibility

Consider the case of an experiment conducted in Morocco between 2006 and 2009, in partnership with the microcredit organization Al Amana. It sought to measure microcredit's impact on household income and consumption in rural areas. It considered pairs of villages exhibiting similar traits: one village per pair was randomly selected to receive microcredit services immediately after an Al Amana agency opened, while the other village served as a control (without receiving the services) for two years. In total, 81 pairs of villages were selected from across Morocco and 6,000 households were included in the study (and surveyed before the agency was established, one and two years later). "Through random assignment, it is possible to constitute a group of recipients and a control group that are initially perfectly similar, not only in terms of statistically observable variables but also in terms of unobservable … variables. Thus any future difference noted between the two groups can be interpreted unambiguously as an effect of the measure being tested" (Bérard & Valdenaire, 2013).[2] This simplicity (i.e., the discrepancy between two group averages) is supposed to make the results

[2] For a presentation of this method, one may consult, in addition to the articles available in la *Vie des idées*, Jatteau (2013a).

unusually legible. Thus “thanks to randomization,” Duflo stresses, “assessing impact is very transparent and simple.”

A Claim to Renew Social Policy by Evidence

The rigor and readability of these randomized experiments make it possible “to revolutionize the social policies of the twenty-first century, just as clinical trials revolutionized medicine in the twentieth century (Duflo et al., 2004). Thus we enter a new age of expertise and public policy, an age of scientific objectivity, innovation, and efficiency, a far cry from the “three I’s—ideology, ignorance, and inertia” that this experimentation would finally cure (Banerjee & Duflo, 2011). According to Duflo, “we must shift away from these big endless debates. Evaluation is rigorous. There is no room for interpretation. Either it works or it doesn’t. If not, one simply has to try something else.” [3] Accordingly, randomized experimentation makes it possible to settle highly ideological debates through rational and dispassionate scientific analysis. It is through the objective, quantified, and incontrovertible measurement of the effects of public policy—through the “pedagogy of evidence” (Jatteau, 2013b)—that the political obstacles and stalemates can be finally overcome.

2. Practical Complications: Contingencies of Protocol, Tricky Results, Bounded Relevance

Even so, the conduct and interpretation of these experiments often prove tricky, to the point that some of the method’s theoretical advantages are, in practice, canceled out.

Restrictive Engineering vs. Social Contingency and Diversity

Carrying out randomized experimentation requires complex engineering and involves diverse groups of actors. As a result, it is exposed to social contingencies that may compromise a protocol’s internal validity.

If social experimentation, compared to laboratory experiments, is a real-life *in vivo* investigational device, it should not be overlooked that protocols produce artificial constraints and varied social reactions: resistance, lack of interest on the part of targeted audiences, the misappropriation of treatment schemes, placebo and nocebo effects, and so on.

For example, the content of an assessed project can be rejected. This was the case of the “classroom cash reward” (*cagnotte scolaire*) mechanism tested in the Créteil (France) school district. Seeking to modify the incentive system for students by introducing financial rewards

[3] http://www.lejdd.fr/Economie/Actualite/Intellectuelle-de-terrain-166936.

designed to encourage attendance, it met with resistance, which resulted in it being canceled at the conclusion of its pilot phase (Bureau et al., 2013). There are also cases of vulnerable populations reacting with disinterest or defiance.

Resistance can also be directed at random sampling itself. Thus two flagship pieces that are constantly cited to illustrate the merits of randomization—particular that of Kremer and Miguel (2004) on the positive impact of deworming pills on the education of children in Kenya—were in fact only quasi-experiments (Deaton, 2010): local partners would not permit the use of random numbers for assignment of schools, hence the use of alphabetization.

These social contingences result in statistical difficulties. Some experiments must deal with low participation or a drop out of subjects over the course of the experimentation process (i.e., attrition), making it difficult to constitute and maintain unbiased samples of an adequate size. Problems relating to group permeability also arise: members of the control group can arrange to get access to treatments, even as, conversely, participants in the treatment group fail to receive them (Devaux-Spatarakis, 2014).

In the effort to control the protocol from beginning to end, experimentation depends on reservoirs of energy and ingenuity. This is because an experimental system requires, if it is to preserve its internal validity, very rigid and highly standardized "treatments." Rigidity and standardization can, however, clash with the flexibility and diversity of the social relations at play in a particular system (for instance, seeking advice on finding a job, mentoring individuals in difficult situations, loans, and so on). This can be seen in a qualitative study conducted alongside the Al Amana experiment. Microcredit is not a mere technical procedure: it is embedded in the social and religious beliefs bound up with credit and debt, it participates in a variety of agro-ecological configurations, and it draws on social interactions with credit officers and local leaders that vary considerably from place to place and even from person to person. Al Amana was, depending on the region, perceived as a branch of the central government—which was seen as frightening in some places, and illegitimate (i.e., "thieves' money") in others—or as a charitable rather than a lending organization (Morvant-Roux et al., 2014). This explains the great variability in impact from region to region and within regions, with the take-up rate varying from 5 to 43% within a single region. The implementation of a program can differ across space as a function of location, values, and routines of the operators, but also across time: programs themselves can evolve over the experimentation process as actors pragmatically take note of certain difficulties (in the case of Al Amana, one sees the abandonment mid-project of a credit quota for women and the introduction of a credit formula for individuals, rather than just for groups, as had been initially planned). Thus it becomes difficult to know what and which target audiences are being tested (Bernard et al, 2012). Mandatory protocol standardization has difficulty adjusting, in this way, to local adaptations and social variability.

This problem also appears in biology, where experiments must confront life's inherent variability. Jean-Paul Gaudillière (2006) reminds us of attempts to standardize the laboratory animals needed to ensure that experiments are reproducible and comparable, notably the physiological and genetic standardization of animal lineages and the creation of homogenous living and feeding conditions in breeding farms. Yet despite such standardization, interactions between handlers and laboratory animals can disrupt experiments: thus the way in which laboratory personnel handle mice (i.e., gently or roughly) has a significant influence on experiment results. This is especially true in clinical trials, which are social constructs exposed to multiple influences (Labrousse, 2010).

For those reasons, one frequently observes major discrepancies between the planned protocol and the way it is implemented on the ground. Upholding a protocol can prove to be an impossible mission. Thus a J-PAL research assistant in Africa mentions experimental practices that resemble ROCT (Randomized Out of Control Trials) (Jatteau 2013b: 20), rather than RCT (Randomized Controlled Trials). In France, a researcher recalls encountering difficulties that were great enough to force him to resort to quasi-experimental techniques: "there are so many adjustments, there are so many patches in it that it is just as contestable, or if it's not contestable, it is hardly convincing" (Devaux-Spatarakis, 2014: 455). When tested on the ground, experiments require numerous expedients that dilute the methodological purity claimed by some "randomistas." Any methodology, whether quantitative or qualitative, entails patchwork. Do these messy tinkering processes not turn this technique into an ordinary tool, requiring additional reflexivity on the part of experimenters?

A Tricky Process of Interpretation

Similarly, the interpretation of results is not as clear-cut and unambiguous as they appear on paper. In the first place, it is difficult to isolate the tested impact. Far from being a straightforward "verdict of the data," identifying what an experiment has actually tested is hardly self-evident (Bernard et al, 2012). In the case of the Al Amana experiments, the outcomes seemed perfectly clear: the rural microcredit program had failed (as evidenced by its very low take-up rate and its insignificant impact on poverty, consumption, and activity diversification). The qualitative study, however, showed that the reimbursement schedule, conceived in an urban environment, did not correspond to the constraints of the agricultural calendar. Thus what is tested is a system that, while apparently simple, represents an array of explicit and implicit claims; yet it is difficult to determine which of these contributed to the experiment's success or failure. This situation exemplifies a well-known epistemological problem: the Duhem-Quine thesis. It is impossible to test a hypothesis in isolation, as any empirical test of this hypothesis (in this instance, is microcredit an effective tool against poverty?) requires one or several auxiliary hypotheses (in this case, the calendar is causally neutral). Experiments seek to isolate pure effects, but isolation often proves ambiguous. The idea of *experimentum crucis*—that is, of experiments that can settle a debate once and for all—seem utopian.

It is equally tricky to grasp the causal path (how? through what mechanisms?) that leads to a particular set of observed results (whether it works or not). Indeed, aside from instances of simple mono-causality (a cause brings about an effect, with no feedback of the effect onto the cause), randomized experiments provide evidence of effectiveness (a particular effect is observed) rather than causality (what mechanisms generated this effect?). Thus clinical trials show that acupuncture is effective in preventing post-operative nausea, but the mechanisms that generate these effects are not known (Labrousse, 2010). In cases of complex, cumulative, multifactorial, and non-linear causality, causal chains become a kind of black box for experimenters. Systematic and complementary qualitative studies should make it possible to open this box.

The Limitation of Relevance to Particular Types of Public Action

As randomized experiments are relatively rigid and suited to a straightforward causality, they prove relevant to projects in which the causal link between the treatment and its effect takes places in a relatively fast and linear way, as in the case of the deworming programs studied by Kremer and Miguel (2004). This simple treatment (one pill every six months) quickly improved the health of children suffering from worms and helped to reduce absenteeism. Yet experiments seem far from adequate, however, when what must be tested is a bundle of complex and evolving measures that depend on lengthy learning processes. Bernard, Delarue & Naudet (2012) have studied these questions in depth at the French Development Agency. They described the projects aligned with randomized experiments as "tunnel projects." In this way, they listed randomization's prerequisites: the tested program must involve "(i) a period that is consistent with the hypothetical causal chain; (ii) a limited number of homogeneous and precise treatments; (iii) an administration procedure that has been tested beforehand; (iv) a causal chain that is brief and independent of external events; (v) adoption of the treatment by the beneficiaries in a way that is quick and stable over time; (vi) participation on the part of the beneficiaries that is broad and stable over time; and (vii) a set of effects measurable over the short and medium term covering the main aspects of the treatment." As many social actions diverge from these prerequisites, experimentation's field of relevance is ultimately very limited. These conclusions reconfirm in a troubling way the literature of the 1970s and 80s that took stock of the previous experimental wave (Monnier, 1992).

3. A Recurrent and Cyclical Project of Governing by Evidence

The practical problems encountered during the experiments are not new. They belong to a long-standing history, that of an "experimenting society"—of the hopes it raised and the (relative) disenchantment that it caused.

From the 1920s to the Present: The "Experimenting Society," a Long-Standing and Recurrent Project

When one expands one's perspective to a longer history of social experimentation, it becomes apparent that psychologists played a pioneering role, well before Fisher's experiments in agronomy in the 1930s and the advent of clinical medical testing in the 1940s. By the 1910s, educational issues and schoolchildren were the privileged subjects of these experiments: "children, like rats, are available in quantity and without cost" and their submission to the authority of teachers lent itself to respect of the experiment's protocols (Boring, 1954, in Dehue, 2001: 290). At this time, American psychological journals published numerous controlled experiments (with test and control groups) dealing with the impact of variation of class sizes, the sex of teachers, and various forms of classroom ventilation. To establish such comparable groups, these psychologists first devised matching methods, before proceeding to randomization. In the mid-1920s, the first major controlled randomized social experiment took place in Chicago. It tested the impact of an information campaign—relating to voting dates and procedures, written in citizens' mother tongues—on voter participation: 6,000 citizens of various origins participated. One can see that the modalities and themes of these pioneering experiments converge with those of contemporary experiments. They both belong to the same project of an experimenting society governed by evidence and they seem to have arisen in similar contexts.

Thus in the United States, a propitious convergence of factors was contributing, by the 1920s and 30s, to the gradual appearance of an experimenting society (Dehue, 2001): first, the rise of social policies and administrative rationalities (and thus of social preoccupations and the model of rational, standardized, and impersonal expertise in governing populations); soon, these would be linked to the tenacious right-wing suspicion of the inefficiency of public funding, which justified the requirement that their usage be objectively evaluated. A similar situation would recur during the second great experimental wave that took place in North America in the 1960s and 70s: social programs such as the Johnson Administration's "war on poverty" led to the rationalization of budgeting systems designed to ensure that public funds were used efficiently (Monnier, 1992). In France, in the mid-2000s, it was also in the twofold context of a need to rationalize budget procedures and a desire to renovate social policies (at the initiative of Martin Hirsch) that experiments were first introduced as part of the third great experimental wave.

Another important factor in the development of experimentation was the mobilization of the economy during the Second World War. It contributed to the development of clinical trials through the mobilization of resources, the availability of numerous subjects, and government coordination of trials. It also fostered psychological experimentation in the American military: the experimental section of the "Morale Division" notably consisted of psychologists charged with evaluating soldiers' motivations and assessing, through randomized experimentation, the impact of the *Why We Fight* film series on these motivations (Dehue, 2001). Donald Campbell, a major figure of the second experimental wave, earned his first stripes as an army psychologist in this division.

Enthusiasm-Deception Cycles in Experimental Evaluations

First introduced in economics in the late 1960s, randomized controlled experiments would experience a significant rise in the 1970s, after which it became mainstream and subsided in the 1980s, before resurging in the 2000s. It seems to follow a cycle: an initial phase of enthusiasm, followed by a phase of relative deception.

In the ascendant phase, that of the "flamboyant experimental paradigm" (Monnier, 1992), major and frequently expensive experiments occur. It is believed that they will significantly reshape social programs as well as the social sciences. Experiments are presented as a depoliticized tool for public action, in which the experimenter is viewed as an impartial expert, an agnostic who relies on nothing but pure facts. Thus in the early 1970s, Campbell saw the social scientific researcher as the "methodological servant of the experimental society," a "scientific, non-dogmatic, honest, and accountable society" (Monnier, 1992).

Then the initial enthusiasm fades. "The [political] sponsors, disappointed at the inability of academics to formulate conclusions in terms of their broader policy implications, abandon the idea of establishing substantial mechanisms narrowly focused on a single goal" (Monnier, 1992: 45). This is the "step-by-step" period, when experiments become less intrusive and more modest: they seek to blend into the ordinary administrative and social operations in order to limit reactions of resistance; they are conducted over shorter periods and require less funding. They test incremental changes within existing systems rather than ambitious new programs (Greenberg et al., 1999). As for experimentation's scientific promoters, they become more measured and cautious. For instance, Cook and Campbell's handbook (1979) raises the number of "potential threats to validity" from twelve to 33 compared to Campbell & Stanley (1963). While one lacks sufficient perspective in the French case, some clues suggest that we are now entering this phase. The number of experiments has dwindled since Martin Hirsch's departure and it seems that government agencies are using the results of these methods infrequently (Devaux-Spatarakis, 2014). Initially presented as a self-sufficient gold standard for assessment, randomized experimentation has become "one tool among others in the evaluator's toolbox" (L'Horty & Petit, 2011). And less intrusive experimentation is increasingly encouraged, in which subjects are not necessarily aware that they are involved in an evaluation.

Even so, randomized experimentation is still continuing its rise in developing countries among international organizations, foundations, and in the academic fields of economics. This latest wave is driven by factors that are idiosyncratic to the discipline of economics: the empirical turn of (some of) the mainstream, increasing connections with experimental psychology, valorization in the most esteemed economic journals, which was not case during the previous wave. Witness the concomitant rise, in conjunction with J-PAL, of the experimental work of John List at the University of Chicago, which is more steeped in behaviorism and neoclassical references than Duflo's.

Repeated Experience of Experimentation: What Lessons?

At times, the history of experimentation seems to stutter. Randomized experimentation made a dramatic entrance into the economic realm in 1968 with the launching of the New Jersey Income Maintenance Experiment at the initiative of an MIT PhD student, Heather Ross (Greenberg et al, 1999). Another PhD student in economics at MIT, Esther Duflo, would play a significant role in creating a second experimental wave thirty years later.

Yet in the literature of the new experimental wave, particularly as it deals with developing countries, little reference is made to the experience of the previous waves—as if its application to new territories was a blank page. Developed in the early twentieth century with the intention of improving social programs in developed countries, the method has, it would seem, returned to its virginal status in the southern countries, legitimating, through a boomerang effect, a powerful resurgence in the North: "One of the essential messages of the MIT professor is that experimental evaluation has proved its worth in analyzing the causes of poverty in poor countries and that it must now be used for the same purpose in rich countries, notably France (L'Horty et Petit, 2011). Within the American Evaluation Association, the protagonists of the previous period were struck by this turn "back to the future." Nick L. Smith, for instance, expressed surprise: "When discussions of the role of experimental design in evaluation became increasingly public a few years ago, I thought, 'Didn't we already settle this?' Almost 25 years ago, I organized and moderated a debate at the 1981 meeting of the predecessor organizations of the American Evaluation Association…: 'Should the federal government mandate the use of experimental methods in evaluation?" (Devaux-Spatarakis, 2014: 109). The association warned against the idea of a gold standard in evaluation and called attention to the limits of randomized experimentation, yet its message finds little resonance among economists.

Yet learning the lessons of past experiments could perhaps accelerate the learning effects among political as well as scientific actors. Randomized experimentation, a useful tool, is neither a gold standard nor a revolutionary method. It would benefit from being integrated into mixed methods approaches (Morvant-Roux et al. 2014; Labrousse, 2016) that would allow for the development of more contextualized usage (for whom? in what context?). The question of the choice of "treatments" and their relevance for populations is as fundamental as that of their efficiency. Rather than viewing it as a depoliticized tool of social action, it must transform itself into an instrument serving democratic debate and turn from "evidence-based policy" to "evidence-informed policy."

Further reading

- Banerjee Abhijit & Esther Duflo (2011), Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty, New York: Public Affairs.

- Bérard Jean & Mathieu Valdenaire (2013), "L'expérimentation pour renouveler les politiques publiques?," La Vie des idées, June 25, 2013, http://www.laviedesidees.fr/L-experimentation-pour-renouveler.html
- Bernard Tanguy, Jocelyne Delarue & Jean-David Naudet (2012), "Impact Evaluations: A Tool for Accountability? Lessons from Experience at Agence Française de Développement," The Journal of Development Effectiveness, vol. 4, n° 2: 314-327.
- Bureau Marie-Christine, Sarfati François, Simha Jules & Tuchszirer Carole (2013), "L'expérimentation dans l'action publique," Travail et Emploi, 135, 41-55.
- Deaton Angus (2010), "Instruments, Randomization, and Learning about Development," Journal of Economic Literature, vol. 48, n° 2: 424-455.
- Dehue Trudie (2001), "Establishing the Experimenting Society: The Historical Origin of Social Experimentation According to the Randomized Controlled Design," American Journal of Psychology, 114-2, p. 283-302.
- Devaux-Spatarakis Agathe (2014), La méthode expérimentale par assignation aléatoire: un instrument de recomposition de l'interaction entre sciences sociales et action publique en France?, Ph.D. thesis in political science, University of Bordeaux, October 2014.
- Duflo Esther, Rachel Glennerster & Michael Kremer (2004), Randomized Evaluations of Interventions in Social Service Delivery, MIT.
- Duflo Esther (2009), Expérience, science et lutte contre la pauvreté, Paris, Fayard.
- Greenberg David, Mark Shroder & Matthew Onstott (1999). "The Social Experiment Market," Journal of Economic Perspectives, 13 (3): 157-172.
- Jatteau Arthur (2013a), Les expérimentations aléatoires en économie, Paris, La Découverte.
- Jatteau Arthur (2013b), "Expérimenter le développement ? Des économistes et leurs terrains", Genèses, 4 (93): 8-28.
- L'Horty Yannick et Pascale Petit (2011), "L'évaluation aléatoire : un succès qui ne doit rien au hasard", La vie des idées.
- Labrousse Agnès (2016), "Not by Technique Alone. A Methodological Comparison of Development Analysis with Esther Duflo and Elinor Ostrom," Journal of Institutional Economics, Issue 12(2), June 2016 (online October 2015), 277-303.
- – (2010), "Nouvelle économie du développement et essais cliniques randomisés: une mise en perspective d'un outil de preuve et de gouvernement," Revue de la régulation, n° 7. http://regulation.revues.org/7818

- Mayneris Florent (2009), “L’économie du développement à l’épreuve du terrain – Entretien avec Esther Duflo,” La vie des idées.
- Monnier Eric (1992), Évaluations de l’action des pouvoirs publics, 2e éd., Paris, Economica.
- Morvant-Roux Solène, Guérin Isabelle, Roesch Marc & Moisseron Jean-Yves. (2014), “Adding Value to Randomization with Qualitative Analysis: The Case of Microcredit in Rural Morocco,” World Development 56, 302–312.

First published in laviedesidees.fr, March 4 2016. Translated from the French by Michael Behrent with the support of the Institut Français

Published in *Books & Ideas*, January 23 2017.