books & ideas.net

Randomized Evaluation: A Success Owing Nothing to Chance

Yannick L'HORTY and Pascale PETIT

Randomized evaluation has become very fashionable. Initially developed in the field of development economics, it has now spread to many public policy areas. Yannick L'Horty and Pascale Petit here discuss the advantages and the limits of this relatively recent tool for evaluating social policies.

The crisis nourishes the development of new social needs for ever increasing numbers of people, even while it limits the capacity of public action to satisfy those needs. This increase in budgetary constraints requires public decision makers to scrutinize the costs and benefits of their policies, and they are now being led to evaluate the impact of spending across the board. The scarcity of public funding makes it more imperative to measure its effects. This pressure to evaluate is particularly noticeable in the social policy field, where expectations are stronger and a large number of measures co-exist, and where these measures are implemented by several actors, at different territorial levels. French contributions to the success of randomized methods of evaluation have occurred in this favourable atmosphere. After having been implemented in several other countries, especially in the USA, Mexico and Canada, these methods are now spreading in France, where they are being applied to hundreds of local social programs.

A Supply Shock and a Public Demand Shock

The way randomized evaluation works is extremely simple. By random selection, you form two groups from a list of the potential beneficiaries of a social program. The test group has access to the program, while the control group does not participate in it. Then you just have to compare the two groups, in order to evaluate the program's results. This procedure makes it possible to resolve quite a few technical problems that arise in the evaluation of public policies (we say more about this below).

However, the procedure involves a temporary departure from the principle of equality; in fact, it was in order to give it a legal basis that the French constitutional reform of 2003, concerning decentralization, was adopted the same year as the basic law on experiments that are carried out by local governments. Social experiments are permissible as long as they have a purpose that is geographically circumscribed and of limited duration, and are conducted with an eye on generalization. Moreover, they must be undertaken at the initiative of a local government, and they must be evaluated. The experiment of the Active Solidarity Income Supplement (RSA[1]), outlined in the 21 August 2007 law on "TEPA" (*travail, emploi, et pouvoir d'achat*: work, employment, and purchasing power), constituted the first large-scale social experiment in France, although in the end it was not evaluated by a purely random method, since neither the beneficiaries of the experimental RSA nor the experiment's test territories were randomly selected.

The impetus given by Martin Hirsch to the development of social experiments has been crucial. It was Hirsch who initiated the RSA program. In June 2007 he became High Commissioner for Active Solidarity against Poverty[2], and in January 2009 he was promoted to High Commissioner for the Young[3]. He has consistently supported the development of social experiments and their evaluation by means of randomized methods. He launched a first call for a social experimentation project in 2007, with a budget of six million euros. This trial scheme was followed in 2009 by a series of calls for projects launched by the Committee on Experiments for the Young (FEJ[4]), which was created by article 25 of the law making the RSA available nationwide on 1 December 2008, with a total budget of 150 million euros, two-thirds from public sources and the rest from the private sector. From this were financed nearly 460 new projects, a large portion of which provided for randomized evaluation.

In addition to the public demand shock, the development of randomized evaluation methods can be explained by another positive shock, this time from the evaluation supply side. Here, the dissemination of Esther Duflo's works played a crucial role in inspiring the enthusiasm of many French researchers. One of the basic messages of this MIT professor, who is one of the founders of the international J-PAL network, is that experimental evaluation has shown its capacity to analyze the causes of poverty in poor countries, and that it can now be used for the same purpose in rich countries, including France. This message was to be a theme of her Inaugural Lecture as the first holder of the Collège de France professorial chair "Knowledge against Poverty"[5] (Duflo 2009), and in her two books intended for a wide readership (Duflo 2010a and 2010b). The theme was then picked up by several French economists, including François Bourguignon, Director of the Paris School of Economics[6] and President of the National Committee for the Evaluation of the RSA Experiments[7]; and Marc Gurgand, who presided over the Scientific Council of the FEJ. And many conferences have publicized the contribution of experimental methods: the social integration meetings launched at Grenoble in November 2007; the conference on "Experiments for Public Policies on Employment and Training," organized in May 2008 by the Employment Ministry's Bureau for Research, Study and Statistics; and the National Conference on Social Experiments, organized by the two High Commissioners – for Active Solidarity and for the Young – in March 2010.

This combination of a change in the institutional framework and these positive shocks of supply and demand explains the development of randomized evaluation in France. For this combination to work, there had to be a pairing up between supply and demand. Now it

[1] *Revenu de solidarité active*

[2] *Haut Commissaire aux Solidarités actives contre la Pauvreté*

[3] *Haut Commissaire à la Jeunesse*

[4] *Fonds d'Expérimentation pour la Jeunesse*

[5] *Savoirs contre pauvreté*

[6] *École d'économie de Paris*

[7] *Comité national d'évaluation des expérimentations du RSA*

happens that the new technologies of evaluation have a specific feature that makes it possible for them effectively to meet the new demand for evaluation. They can be applied in a context in which the data are rare, or even non-existent, so that data are actually constructed by the evaluation. While the econometrics of quasi-experimental evaluations requires pre-existing large data bases in which the number of observations are counted by the tens of thousands, randomized evaluation methods make it possible to discern the effects of a program in which the beneficiaries are counted only by the hundreds. This exactly corresponds to the limits of the local micro-programs for which there has been such a rapidly expanding demand in the French context of the increasing decentralization and territorialisation of public actions.

The Good Points of Randomized Methods

Behind these contextual facts explaining the success of randomized evaluation methods are their intrinsic positive characteristics. Whenever it is a question of quantifying a social program's impact (*ceteris paribus*) on a range of examined variables, randomized methods offer full satisfaction of the evaluators' expectations. Indeed, random selection provides an excellent reply to what James Heckman (1992) calls "the problem of evaluation". If you want to measure a program's effects, it is important to be able to observe an individual in two different situations, one in which he or she benefits from the program, and another one in which the same individual does not benefit from it. But how can you make these observations, given that for every individual, only one state of things is actual? How can you know what the behaviour of a program beneficiary would have been if he or she had not had access to the program? Responding to these questions amounts to being able to establish a control group – also referred to as the "counterfactuals" – who at best mimic the behaviour of the members of the test group. The whole problem of evaluation is to construct a good control group. In randomized evaluations, these counterfactuals have a special status. They do not appear fortuitously by nature, as in the case of natural experiences. They are constructed *ex nihilo* by the researcher. Yet they are not invented, but observed.

Having high-quality counterfactuals avoids the risk of selection bias. The problem is in monitoring the heterogeneity of both the observable and the non-observable variables. To achieve this without changing with the experimental framework, it is necessary to use a very sophisticated method of data analysis, such as the matching method suggested by Donald Rubin, which requires a large number of observations. But changing to using an experimental framework of randomized evaluation guarantees that the persons in the two groups will have on average the same observable and non-observable characteristics. If the two groups are big enough (i.e. if they contain several hundred individuals), they will be identically structured by age, sex, educational level, and also by other characteristics that cannot be observed but which can affect the variables being examined – for example, motivation, preferences, cognitive capacities, etc. The main advantage of a randomized selection is that it avoids selection biases, thus producing a robust quantification, with a great economy of statistical and econometric means.

Beyond this frequently highlighted methodological advantage, randomized evaluation methods change for the better the job of the researchers who are applying them. In Esther Duflo's books and articles, she emphasizes an unexpected advantage of these methods. Because they involve a close partnership between the experimenter, who is bringing changes to public policy, and the evaluator, who is measuring the results – a partnership that is established before, during and after the implementation of the policy – there ensues a special relationship between the two. The experimenter progressively takes on board the requirements of the evaluation, and the evaluator gradually becomes a co-constructor of the policy being evaluated. The researchers gain an intimate knowledge of the policy being implemented, which makes it possible for them to be ever more creative in their investigations. Thus the quality of their research is improved.

One could add that the researchers doing randomized evaluations significantly increase their grip on their research purposes and methods. Not only can they become co-constructors of the public policy, they are also no longer so passive with regard to the data, which they construct or completely reconstruct as appropriate. Applied economics becomes field research. The researchers spend less time in their offices working on models or programs, and more *in situ* implementing experimental protocols. Because of this greater involvement of the researcher, the experimental evaluation becomes more transparent for the experimenter and for the financer. It becomes more understandable by the policy makers, who will often find it more credible than traditional approaches that do not have these properties.

There are Two Sides to Every Coin

In our introduction to randomized evaluation methods (L'Horty and Petit 2011), we point out that every advantage accorded to these methods has a downside. First of all, randomization can bring in some problems (overlapping, but not completely) of acceptability and ethics. A randomized evaluation protocol involves depriving some people of resources that might be needed to improve their situation. If these resources do have an effect on people's lives, depriving the control group of them can go against the very purpose of the experimenting institution. After all, the first aim of this kind of institution is to improve the welfare of people with difficulties in social and/or economic inclusion. Acceptability is an issue raised by social workers who are in direct contact with people or elected officials who naturally want immediate access to a social innovation to be available to the greatest number of beneficiaries.

Moreover, an experiment conducted on a specific territory by a specific experimenter might well produce specific, non-generalizable results. In particular, since the experimenters who are supporters of the project are committed to its concept, and determined that it succeed, they will favour this kind of protocol and they will be ready to make the efforts needed to implement it. But their desire to see its success can lead them, however consciously or unconsciously, to introduce into the test group supportive measures that are not included in the protocol. This intervention can influence the measurement of the effects on the test group of the program being tested.

Thus, even if randomized evaluation can prevent selection bias, it displays other kinds of bias. Because the experiments are circumscribed in both space and time, generalizing the program involves changing the temporal and spatial scales, which can produce changes in the effects of the program. Here we are referring to "equilibrium effects". Extending a local mechanism has aggregate effects that modify market equilibriums and prices. Through these effects, the program has an impact on the control group that is not considered in the randomized evaluation. So the problem is how to know whether the result observed at the local level will be the same when the experiment is made general and the aggregate effects come into play. For example, Rodrik (2008) argues that supplying free anti-malaria mosquito nets, even if effective in a randomized experiment, cannot be instituted on a country-wide scale, because that would dry up the market distribution networks that supply rural areas.

One other noteworthy kind of obstacle to the implementation of experiments and to their evaluation by demanding protocols concerns the material difficulties encountered by the executing agents and evaluators of these programs. Because a randomized evaluation involves checking up on the experiment before, during and after its execution, it entails logistical costs that are often very high, and production delays that are sometimes substantial. That is why the timing of the evaluation does not always relate to the timing of the public decision. What is more, the evaluators will always be confronted with a great many things that are unforeseen and unforeseeable. They think they are evaluating an experiment, while in reality they are experiencing an evaluation.

An Additional Tool for Evaluators to Use

It is clear that the development of randomized evaluations undeniably constitutes progress, enlarging the panoply of tools for researchers interested in social issues, but it is not leading to a new methodological gold standard that is going to displace all the other approaches. Experimental methods can be a tool of choice in the ensemble of evaluation methods, but their range and their limitations must be kept in mind, in order to understand this tool and to use it wisely.

Not all public policies are subject to randomized evaluation. Macroeconomic, monetary and budgetary policies, as well as broad structural activities in the fields of taxation, social protection, and even industrial and commercial policies, cannot be evaluated by this kind of method. Just because they cannot be evaluated does not mean that they should not be implemented. In the panoply of evaluation tools, there is a place to be taken by randomized methods. These methods should not take all of the places, but they should take their place.

So as not to be misunderstood: having a nuanced view about the benefits and limits of these methods should not encourage inertness. Although experimental methods have been used for a long time in the hard sciences, in medicine and agronomy, and even in marketing, and since the 1960s have been used to evaluate big social programs in rich countries, mainly in North America, their introduction has been very tardy in France, which has some catching up to do. In fact in many cases randomized evaluation is the only quantitative approach that is suitable for a local social program that targets some hundreds of beneficiaries. Not deploying this kind of methodology here amounts to renouncing any quantifying evaluation of the effects of social policies, which is probably the worst outcome from the point of view of public decision making.

References

Deaton A., 2009, “Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development*”,* NBER Working Paper, no. 14690, January.

Duflo E., 2009, “Experiments, Science and the fight against Poverty”, Inaugural Lecture, Chair in *Savoir contre pauvreté*, Collège de France, Paris, 8 January.

Duflo E., 2010a, *Le développement humain: Lutter contre la pauvreté (I)*, Le Seuil, “République des idées”, Paris, 104 pp.

Duflo E., 2010b, *La politique de l'autonomie: Lutter contre la pauvreté (II)*, Le Seuil, "République des idées", Paris, 104 pp.

Heckman J., 1992, "Randomization and Social Policy Evaluation," in C. Manski and I. Garfinkel (eds.), *Evaluating Welfare and Training Programs,* Cambridge, MA: Harvard University Press.

L'Horty Y. and Petit P., 2011, "Évaluation aléatoire et expérimentations sociales", *Revue Française d'Economie,* VOL XXVI, juillet, pp 13-48

Rodrik D., 2008, "The New Development Economics: We Shall Experiment, but How Shall We Learn?" HKS Faculty Research Working Paper, no. 08-055.

Published in Books&Ideas, on December 12th 2013. Translation by John Zvesper, with the support of the Institut du Monde Contemporain.

First published in French on laviedesidees.fr, 11 May 2011.