3rd International Conference

on Public Policy (ICPP3)

June 28-30, 2017 – Singapore

Panel T07P03 Session ?

Expertise and Evidence in Public Policy

Title of the paper

Experiment-based policy making in France: political use of science and practices-based knowledges

Author(s)

Agathe Devaux-Spatarakis, Centre Emile Durkheim/Quadrant Conseil, France, adevaux@quadrant-conseil.fr

Date of presentation

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Abstract

The French government opted for a model of Evidence-based policy in favor of new evidence produced by the evaluation of pilot programs by randomized controlled trials organized by the Experimental Fund for Youth created in 2009.

The thorough empirical study of the learning interests of the various stakeholders of these experiments, reveals potential conflicts during their design, their implementation and use of the evidence produced. Conducted for the sake of producing sound evidence for future policy program, these experiments were mainly used by project managers to produce practical knowledge, and latter subject to political misuse of results information regardless of the quality of the evidence produced.

Keywords

Experiments, political misuse, conceptual learning, practical knowledge, Evidence-based policy

The beginning of the new century witnessed a worldwide growing interest for impact assessment by evaluation methods, and a general call for Evidence-Based Policy (hereafter EBP) within international and national organizations (Donaldson, Christie, & Mark, 2009; Evaluation Gap Working Group, 2006). Evidence-based Policy movements call for policy initiatives to be more systematically supported by evidence and they organize the encounter of a supply and demand for evidence (Sanderson, 2002). In doing so, the strategies of these movements are twofold. First, they intend to act upon the supply side of evidence by improving the quality of studies produced, and second, they attempt to foster the use of evidence by policy actors, by institutionalizing a more systematic use of evidence in policy making (Lee, 2004; Parsons, 2002). An overview of EBP movements around the world shows that the strategies of EBP movements vary depending on the characteristics of national evaluation suppliers, national institutions, and political culture (Rieper, 2009). Across countries, a diversity of methods for impact evaluation are promoted, different types of suppliers of evaluation are identified, and various types of knowledge synthesis are delivered to policy makers.

The model introduced as the most scientifically based, following the standards of evidence-based medicine practice, was evidence informed by experiments – or pilot programs – assessed by Randomize Controlled Trials (RCTs)(Cochrane 2011; Coalition for Evidence-Based Policy 2007). This trend of the EBP movement, promoted a new kind of scientific legitimacy to advise policy, based on the empirical demonstration of pilot program's impact, rather than grounded on scientific general expertise on the policy area (Duflo 2005). In a nutshell, effectiveness of new policies had to be demonstrated through pilot programs assessed by controfactual analysis before generalizing the intervention to the whole population.

This model of EBP made its way to France through the creation of the Experimental Fund for Youth (EFY) within the French government, in 2009. Its ambition was to design an array of new policies for the French disadvantaged youth, grounded on sound evidence from funded pilot programs, evaluated preferably by RCTs (Conseil Scientifique du FEJ 2009). This new organization dedicated to EBP begs the questions of what was effectively learned by experimenting and most importantly how was this knowledge effectively utilized to fuel policy making?

Literature has extensively addressed scientific debate on the most relevant methods to produce evidence or to identify barriers and facilitators of the use of evidence by policy makers (Donaldson et al., 2009; Duflo, Glennerster, & Kremer, 2004; Sanderson, 2002; Solesbury, 2001). However, little attention has been given to the study of the interaction it organizes between the stakeholders involved, and the interests they pursue in conducting or funding experiments. This work wishes to demonstrate that studying RCTs as a social institution is fruitful to understand how its stakeholders interplay with each others and what kind of knowledge is produced. From a constructivist perspective RCTs are a set of rules, resources and roles that are shaped in turn by the reactions of those who are subjected to RCT practices rules.

In order to pursue this inquiry, 50 interviews were conducted among actors from the supply and demand side of evidence in France, as well as a thorough analysis of intern documents produced by the EFY, completed by a cross analysis of 15 embedded cases studies of pilot programs evaluated by RCTs. This research opted for a large definition of learning, thus scrutinizing both scientific knowledge but also practical knowledge from policy managers

implementing the program (Head, 2008; Patton, 2010). Also, all type of uses were under our attention, following C. Weiss typology of political use of evaluation (Weiss, 1998).

This paper will separately address the interests pursed by each group of stakeholders of these experiments:

- Evaluators of the projects, here mainly researchers;
- Project managers, the organization which designed the projects and would implement it;
- Sponsors, the public or private actors funding experiments and their evaluation.

Suppliers of RCTs: academic interest as main motivation

Before studying the use of impact evaluation results by program managers and policy makers, one must account for the nature of the evidence produced by evaluation suppliers . Although suppliers of evidence they were pursuing their own interests in conducting RCTs. First, as stated in publications and confirmed in our interviews, a key interest in this group was to attempt this evaluation method and learn about its implementation (Fougère, 2000; L'Horty & Petit, 2010). Since 2000, the rise of international academics' interest for impact evaluation among micro-economists is documented (Card, DellaVigna, & Malmendier, 2011; Levitt & List, 2008). One of these researchers explains: *« This method (RCTs) has become the norm in applied economics (…) If you want credible results, it is expected to give evidence by this method ».* French researchers not only wished to conduct RCTs themselves, but also to study behavioral mechanisms in order to bridge some knowledge gaps on special issues debated academically. During interviews, they explained how they picked pilot programs depending on the research question they could address it with.

A couple of public servants working for the experimental fund complained about the fact that some scientists seemed determined to stick with RCTs even if implementing conditions turned out to be unfavorable "*With some researchers, we were under the impression that applying the method (RCTs) was the most important [sic]. Whereas the aim of this call for projects was not to conduct pure scientific work but to improve social policies*".

Researchers eventually had to lower their expectations about the quality of evidence produced in these contexts. From the start of the experiment, some projects were not fit to be evaluated by RCTs. They were either too small to gather enough beneficiaries to reach statistical significance, too complex to be comprehensively evaluated by this method, or did not allow strict randomization. Moreover, even projects more adapted to RCTs at the onset could become unsuitable because of low adoption (take up) or modification of the intervention during implementation, both conditions strongly hindering proper implementation of RCT evaluation protocols.

As a result, few experiments met the criteria to be considered as credible evidence by the scientific community. Almost all the results had to be strengthened by additional statistical analysis, or new data collection to increase their internal validity. In the end, no experiment, under this study, led to clear-cut recommendation towards generalization of the pilot program, and evaluators often called for more experiments to draw definitive conclusions. The interests of these scientists were focused on learning about behavioral mechanisms common to many programs from long term cumulative evidence from evaluating diverse programs, and not on improving the approach of one particular program. One of them explained this in an interview as follows: *"This is not an evaluation method shaped to assess if a program works or not. It is an evaluation to answer the question what should we do? What is the problem? We have an interest in mechanisms, in testing hypotheses".*

These experiences allowed researchers to better define conditions to produce evidence with RCTs. It generally appeared that the best evidence was produced when evaluators actively contributed to the design of the project. When collaborating with project managers, evaluators managed not only to establish implementation control to reduce bias, but also to shape the intervention itself, in order to better address a scientific question. Some experiments blurred the lines between the supply and demand side of evidence. Evaluators sometimes solicited project managers in order to find a field for experiments on a mechanism of interests for scientific debate or at least they co-constructed the experiment with project managers. According to our analysis of interviews with evaluators, their ideal project with respect to evaluation was a very simple and punctual intervention, mobilizing one simple technical mechanism, with beneficiaries easy to reach and to monitor , with low media and political interests, and with low stakes for project managers. Commenting on such conditions a researcher explained: *"Here we have the opportunity to experiment, not just to be evaluator, but to choose things, to set variables, to interact with people in the field implementing. We are more proactive, more beforehand".*

As a result, from the supply side perspective of researchers implementing RCTs, the best evidence was produced when projects strictly followed their protocols, with no adaptation of the program along the time period of the experiment. Yet, these requirements overlooked the policy environment of "social experiments", and the particular conditions and interests for learning amongst the demand side. Studying the "demand side" of impact evaluations entails first defining what groups of stakeholders constitute this demand. In the RCTs under study, two groups with different expectations emerged, first program managers conducting pilot program, second policy makers or sponsors funding experiments.

The incompatibility between program managers' learning needs and RCTs

The expectations of program managers leading them to design pilot program to be evaluated by RCTs were manifold. Many project managers viewed the experimental fund as an opportunity to find subsidies to carry-out innovation on the field in times of budgetary cuts. Their main interest for collaborating with RCT evaluators was to increase their chances to be selected for project funding. One of the members of the experimental fund recalled this as follows: *"It was the price to pay to prove that they were right, to fund their project. (…) it was a blessing in disguise"*. The impact of these subsides varied depending on the size of the organization, but to some small NGOs it amounted to more than half of their overall budget. Project managers also generally had two expectations of RCT evaluation results: to gain insights on their project to improve its design and implementation in the field, and to prove their project's impacts in order to secure sustainable funding for larger-scale implementation after the experimental phase was over.

Eventually, along the experiment process, their learning expectations and needs diverged from the evaluation design. When project managers started collaborating with evaluators, they often had not clearly understood the requirements of RCT evaluation protocols regarding project design. Many pilot projects were therefore too complex to be rigorously assessed by RCTs, as they were tailored to offer comprehensive solutions to disadvantage youth. They involved many stakeholders, and were dedicated to hard to reach beneficiaries. Also, they were not strictly planned since project managers were used – and expected – to be able to adjust innovation "as they go" in accordance with the obstacles or opportunities encountered during implementation.

Therefore, rather than complying with RCT protocols they often chose to provide access to the treatment to all eligible members of the public and exercised flexibility in treatment of individual cases. This approach, while ethical, undermined the scientific validity of the experiment. Another dimension of conflict between RCT principles and the reality of field practice had to do with control of treatments. Although most interventions were properly framed at the outset many either evolved as the experiment took place or were changed to meet the distinctive needs of the agents who delivered the intervention. In a nutshell, for a wide variety of reasons, none of experiments examined by this research followed the scientific protocols planned by the academic evaluation team. The discrepancy between experimental protocols and the ways experiments actually unfold in the field has been underlined in the literature even for the most celebrated examples in the use of RCTs for social research[1].

Project managers were accustomed to proceed through learning by doing and therefore adapted the pilot program while it was in its experimental phase in order to improve its impact. Hence, in many cases, and as other observers noted in other contexts, pilot projects were only becoming stable by the end of the experiment (Patton, 2010). This led many project managers to discredit RCTs results as they only focused on how the project was originally designed and ignored many of its modifications. Some evaluators acknowledged this disconnect: *"What were they (project managers) actually doing in the field? We've never managed to figure it out"*. Project managers expected a comprehensive contextualized

[1] FAULKNER, W., A critical analysis of a randomized controlled trial evaluation in Mexico: Norm, mistake or exemplar?, *Evaluation*, 2014, 20(2) 230–243;DEATON, A .,« Instruments, Randomization, and Learning about development », 2010, 48 *Journal of Economic Literature* 424-455

assessment of their pilot projects, whereas the impact evaluations focused on the one dimension of the project that could be best monitored and randomized.

The timeline of the production of impact evaluation results was also misaligned with the program managers' needs. Evaluators issued evaluation results at the earliest 6 month after the end of the pilot program. In some experiments, given the various breaches of RCTs protocols and the time it took to come up with remedial analysis, it took 1 or 2 years. In contrast to the supply timeline, pilot projects demanded evidence much earlier. They were funded only temporally by the experimental fund, and no learning management system had been defined on how to proceed with evaluation results. Consequently, program managers whose organization partly depended on these subsides expected evidence before the end of the experiment in order to solicit sponsors to secure future funding. This led them to only use intermediary results, when those were in favor of the project, or to even conduct surveys on their own to gather the data they considered necessary. They also realized that they had unrealistic learning expectation from RCTs as they were expecting to get a comprehensive understanding of the programme in context, whereas RCT results were only focused on one aspect of the intervention that could be accurately assessed by this method (Rodrik, 2008). As a result, the decision to generalize pilot programs was often not grounded on evidence from impact evaluations, but rather depended on political windows of opportunity or the political influence yielded by project managers. The scarce use of impact evaluation results by program managers may also be explained by the fact that evaluations were funded and imposed upon them by the experimental fund. Therefore, our attention should now turn towards the political committee of the experimental fund as well as other political actors influencing scaling up of pilot programs.

A cherry picking use by political actors regardless of the quality of evidence produced

Studying evaluation use often entails turning the focus of analysis towards institutions or political actors who decided in the beginning to fund these pilot projects and their evaluation. Yet, in the cases under study, political turnover hindered this inquiry. By the time impact evaluation results were issued, the Experimental Fund for Youth had no influence nor subsides to decide on scaling up. Although results of experiments were made public, it remained unclear which institution – and how – should take them on board. No systematic learning management system was established and only a few RCTs results were used at all. To proceed to this analysis, we conducted follow-up studies of all pilot projects, to trace if it was scaled-up, scaled-out or just if the pilot project was maintained.

Overall, it appeared that political interest for impact evaluation had been highly opportunistic. As underlined earlier, one of the main interests were to fund experiments in order to prove efficiency of unpopular reforms. Hence, considerable attention was given to evaluation results likely to back political agendas. Also, political actors rarely stayed in office in the same position more than a few years. Pilot projects can hence move away from being at the center of attention when new elections shift interest to other issues. Although these characteristics of the political arena are well known, looking at practical cases exemplifies how the quality of evidence produced had little influence on actual decisions taken.

First, many local experiments were rolled out nationally before their impact evaluation results were published. Pilot programs were viewed as innovative interventions, to test different ways of implementing a program rather than assessing their impacts. Consequently, some pilot programs were rolled out nationally after only one year of the experimental phase,

provided that they aligned with the political agenda, and its project managers yielded enough political influence. This decision sometimes referred to intermediary evaluation results, but in other cases it did not. Given the breach in evaluation protocols and the difficulties of evaluators to deliver clear cut answers on the overall impact of the interventions, politicians could cherry pick figures that would back their reforms.

Three impact evaluation results were explicitly used to back political decision making at a national level and were displayed as examples of good practice of evidence-based policy making. A closer look at the kinds of "evidence" that was used and who used them lead us to mitigate this judgment.

The first case in a local pilot program aimed at getting parents involved in their child's schooling. The decision to nationally roll-out the pilot project was grounded on positive intermediary results only after one year of experimentation out of two (Ecole d'Economie de Paris, 2011). This decision to fund the program nation-wide was taken by a high civil servant within the Ministry of Education, who happened to be the former project manager of this pilot experiment. Hence, evidence was not consolidated yet, and we can infer that the new high position of the former project manager of this pilot greatly favor this program rolled-out.

The second case is an experiment on the use of anonymous CVs funded by the French government and conducted by the national employment agency. This pilot aimed to test a measure passed by a bill 5 years before by the former government but not yet implemented. Results from this experiments surprisingly showed that the use of anonymous CVs increased discriminations. Indeed, researchers who conducted this evaluation questioned the validity of the negative results of this pilot program on fighting discrimination, as companies participating in the study were aware that they were tested, and therefore subject to the

Hawthorne effect (Behaghel, Crépon, & Le Barbanchon, 2011). Yet , the French president at the time, N. Sarkozy, used these results of an RCT to terminate this program and to back his decision not to implement this bill. Hence, one can make the hypothesis that these negative results, although questionable, were welcomed as useful "neutral" argument against this bill, and subject to political misuse.

More interestingly, the third case, was a pilot consisting in organizing a boarding school for disadvantaged students. Two opposite decisions emerged from this experiment assessed by RCTs. First, although no results were available yet, the president at the time, N. Sarkozy, decided to roll out the experiment stating that it was experimented without any scientific results produced (Cour des comptes, 2014). Second, after the new presidential elections, F. Holland, terminated this program, stating that it was inefficient, although the publication of the final results was not yet issued (Behaghel, De Chaisemartin, Charpentier, & Gurgand, 2013; Beyer, 2013).

These three cases exemplify the weaknesses of an EBP model, driven by suppliers of evidence, and relying only on the strength of RCTs to foster EBP amongst policy actors. The lack of genuine demand for evidence, understanding what kind of evidence can be produced by RCTs, led in France to unmeet learning expectations from project managers, and misuse or non-use of evidence produced by RCTs by political actors in power.

Conclusion

All evidence-based policy approaches tend to bridge a gap between the distinct fields of science, policy and politics through the organization of knowledge transfers. These different approach to knowledge are particularly manifest in the organization of "experimental-based" policy learning. This disconnect calls for consolidating EBP approaches taking into account

these discrepancies, and build rather on compromises between these types of learning than on an attempt to convert one field to the learning processes and interests of the others.

Bibliography

Behaghel, L., Crépon, B., & Le Barbanchon, T. (2011). *Evaluation de l'impact du CV anonyme, synthèse et rapport final* (p. 97). PSE,CREST,J-Pal.

Behaghel, L., De Chaisemartin, C., Charpentier, A., & Gurgand, M. (2013, avril). Internats d'excellence: les enseignements de Sourdun. J-PAL Europe, Institut des Politiques Publiques.

Beyer, C. (2013, avril 11). Vers la fin des internats d'excellence. Consulté 26 mars 2014, à l'adresse http://www.lefigaro.fr/actualite-france/2013/04/11/01016-20130411ARTFIG00350-vers-la-fin-des-internats-d-excellence.php

Card, D., DellaVigna, S., & Malmendier, U. (2011). The Role of Theory in Field Experiments. *Journal of Economic Perspectives*, *25*(3), 39-62.

Coalition for Evidence-Based Policy. (2007). Hierarchy of Study Designs for Evaluating the Effectiveness of STEM Education Project or Practice. Consulté à l'adresse http://coalition4evidence.org/wp-content/uploads/2009/05/study-design-hierarchy-6-4-09.pdf

Conseil Scientifique du FEJ. (2009). *Guide méthodologique pour l'évaluation des expérimentations sociales à l'intention des porteurs de projets* (p. 22).

Cour des comptes. (2014, février). Rapport public annuel, Des internats d'excellence à ceux de la réussite: la conduite chaotique d'une politique éducative et sociale.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Éd.). (2009). *What counts as credible evidence in applied research and evaluation practice?* Los Angeles: SAGE.

Duflo, E. (2005). Evaluer l'impact des programmes d'aide au développement: le rôle des évaluations par assignation aléatoire. *Revue d'économie du développement*, *19*(2), pp 185-226.

Duflo, E., Glennerster, R., & Kremer, M. (2004). Randomized Evaluations of Interventions in Social Service Delivery. MIT. Consulté à l'adresse http://stuff.mit.edu/afs/athena/course/other/d-lab/DlabIII06/random-eval.pdf

Ecole d'Economie de Paris. (2011). *Quels effets attendre d'une politique d'implication des parents dans les collèges, évaluation de l'impact de la Mallette des parents*. Fonds d'expérimentations pour la jeunesse.

Evaluation Gap Working Group. (2006). *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. Center for Global Development.

Fougère, D. (2000). Expérimenter pour évaluer les politiques d'aide à l'emploi: les exemples anglo-saxons et nord-européens. *Revue Française des Affaires Sociales*.

Head, B. W. (2008). Three Lenses of Evidence-Based Policy. *Australian Journal of Public Administration*, *67*(1), 1-11. https://doi.org/10.1111/j.1467-8500.2007.00564.x

Higgins, J., & Green, S. (Éd.). (2011). Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration.

Lee, J. (2004). Is Evidence-Based Government Possible?

Levitt, S., & List, J. (2008). Field Experiments in Economics: the Past, the Present, and the Future. National Bureau of Economic Research, Cambridge MA.

L'Horty, Y., & Petit, P. (2010). *Evaluation aléatoire et expérimentations sociales* (Document de travail No. 135). Centre d'études de l'emploi.

Parsons, W. (2002). From Muddling Through to Muddling Up - Evidence Based Policy Making and the Modernisation of British Government. *Public Policy and Administration*, *17*(3), 43-60.

Patton, M. Q. (2010). *Developmental evaluation : applying complexity concepts to enhance innovation and use*. New York: Guilford Press.

Rieper, H. (2009). The evidence movement, the development and consequences of methodologies in review practices. *Evaluation*, *15*, 141-163.

Rodrik, D. (2008). The new development economics: We shall experiment, but how shall we learn? Présenté à Brookings Development Conference.

Sanderson, I. (2002). Evaluation, Policy Learning and Evidence-based Policy Making. *Public Administration*, *80*(1), 1-22.

Solesbury, W. (2001, octobre). Evidence Based Policy: Whence it Came and Where it's Going. ESRC UK Centre for Evidence Based Policy and Practice, Queen Mary, University of London. Consulté à l'adresse http://www.kcl.ac.uk/content/1/c6/03/45/84/wp1.pdf

Weiss, C. H. (1998). Have We Learned Anything New About the Use of Evaluation? *American Journal of Evaluation*, *19*(1), 21-33.