conseil d'analyse économique
French Council of Economic Analysis

Members of the Conseil d'analyse économique

Public Policy Evaluation

Les notes du conseil d'analyse économique, no 1, February 2013

Public policy evaluation is a difficult exercise, both technically and institutionally. Technically, because a number of pitfalls lie in wait for the evaluator: correlation (between a policy and its results) does not mean causality and the evaluator must take into account reverse causalities and interactions between the policy under consideration and multiple other factors; they must also be aware of the fact that the ultimate beneficiary of a measure is not necessarily the person targeted, and that a policy may have a number of effects that are sometimes far removed from the field targeted initially. A number of statistical techniques make it possible to work around these issues, the key being to be able to reconstruct what would have happened had the policy in question not come into being. Where genuine experimentation is not possible, researchers make use of existing discontinuities in public policy, whether the policy is implemented in successive stages, or it is applied with thresholds (in this case individuals or businesses on each side of the threshold are compared).

Evaluation is also difficult to implement institutionally since only a thorough protocol, defined where possible prior to implementation of the policy, enables a credible evaluation to be made. This protocol must ensure the independence of the evaluators and their access to the data required for the evaluation. It must also make provision for a period of open discussion of the hypotheses and results, within an interdisciplinary framework. Finally, it must leave the evaluators free to publish their results and consult with other experts both in France and abroad. In practice, policy evaluation should not be carried out by the administration tasked with its implementation. Administrative expertise is a vital addition to technical expertise, particularly in order to understand the modes of application of a policy and its interactions with other measures. It must be combined with technical expertise but cannot be a substitute for it. External evaluators must be appointed through a transparent process that is also external to the administration in charge, ensuring the prevention of any relationship of dependency with the commissioner and the promotion of a plurality of approaches. For their part, evaluators must comply strictly with data confidentiality and be utterly transparent about any potential conflicts of interest. Finally, a credible evaluation should be based on the triptych of a coordinator (Parliament, auditing court, etc.), the administrations concerned and independent experts. These elements are not beyond the capabilities of a government that is determined to sift through its public policies.

Although a credible evaluation does take time, a reliable and independent examination means that time-savings can subsequently be made during the decision-making process.

Introduction

The Modernisation of Public Policy announced on 18 December 2012 stipulates that "all public policies, throughout the five year period, shall be evaluated".[1] In fact, the accumulation of measures over the decades means that public policy is not easy to understand at present and probably conceals policies that are obsolete (the initial objectives have been attained), inefficient (the objectives are poorly attained or attained at too great a cost), or misdirected (in practice serving other purposes than those designated). Overall this is costly for public finances and lacks democratic transparency. It is therefore legitimate to seek to evaluate each policy on a case by case basis.

The evaluation of public policies is a difficult exercise: a number of pitfalls lie in wait for evaluators, which may skew and undermine the credibility of an evaluation that does not comply with a strict protocol. However, sound evaluation is not beyond the capabilities of a government that is determined to sift through its public policies. After presenting the classic pitfalls in evaluation, we shall set out methods that make it possible to obtain a credible evaluation of public policies, setting out the requirements, specifically in terms of statistical data. Finally, we shall present the hallmarks of a sound evaluation, which must combine the various levels of expertise in the evaluation protocols to ensure independence and a plurality of evaluators, and dissemination and discussion of their hypotheses and results.

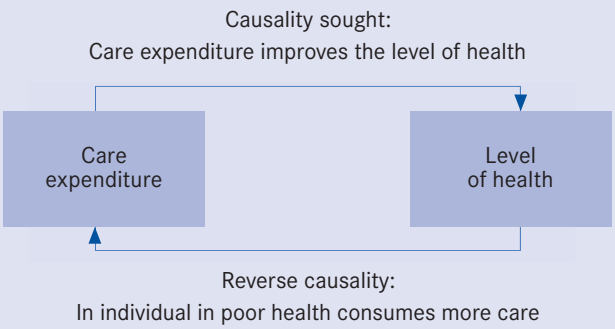
The pitfalls of evaluation

In order to evaluate a public policy, it is not sufficient merely to observe trends in the key indicators targeted by the policy. Below we set out the classic pitfalls in evaluation.

Determining the causal impact of the policy

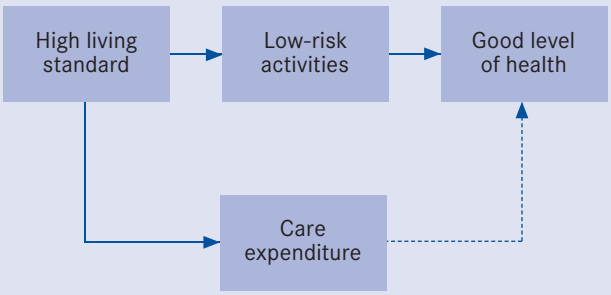
The first evaluation difficulty lies in determining a causal relationship between a policy and a result. Let us suppose, for example, that we wished to evaluate the impact of healthcare expenditure on the level of health of a population. The simple correlation, within the population, between healthcare expenditure and level of health is negative, since the individuals who spend the most are generally the least healthy. This is a case of reverse causality, from level of health to expenditure, which does not provide us with any information on the impact of expenditure (Figure 1).

1. An example of reverse causality



The identification of a causal relationship between healthcare expenditure and level of health also comes up against interference from external factors, such as living standard, which have an influence on both expenditure and level of health: comfortably off individuals spend more on their health and are generally speaking more healthy, amongst other reasons because they carry out lower risk activities. The correlation between healthcare expenditure and level of health does not therefore correspond to any causal relationship (Figure 2).

2. An example of correlation without causality



In order to evaluate a public policy, it is not sufficient merely to observe trends in the key indicators targeted by the policy.

The authors would like to thank Clément Carbonnier who monitored this work within the permanent unit of the CAE.

[1] Declaration of Jean-Marc Ayrault to the Inter-ministerial Committee for the Modernisation of Public Policy, 18 December 2012, available at http:/.www.gouvernement.fr/premier-ministre/declaration-de-jean-marc-ayrault-au-comite-interministeriel-pour-la-modernisation-d

In order to identify the impact of healthcare expenditure on the level of health of individuals, we would need to compare not the level of health of the "major" consumers of healthcare with that of "minor" consumers, but the level of health of the same person individually depending on their consumption of healthcare. Since the same individual cannot consume both a high and a low level of healthcare, our approach must be based on a large number of individuals in respect of whom all of the characteristics liable to influence their level of health can be closely monitored independently of healthcare expenditure. In estimating healthcare expenditure and requirements, Martin et *al*. (2008)[2] were able to demonstrate a positive causal impact: increasing healthcare expenditure to fight against cancer and cardiovascular disease saves lives. Estimates reveal that on average cancer care can add one year of life for 13,100 pounds sterling and for cardiovascular disease one year can be added for 8,000 pounds.

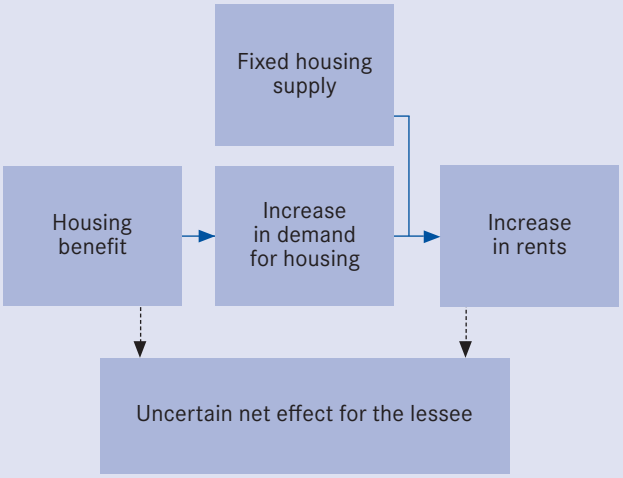
The same types of problems are encountered when we seek to evaluate return to work accompaniment policies. Unless caution is exercised, we may find that those who received accompaniment took longer than other unemployed persons to find a job. But we must take into account the fact that the Job Centre personnel do not necessarily allocate this accompaniment randomly: they may direct accompaniment to those individuals most disadvantaged in terms of employability or inversely, upon those persons who are the closest to employment (particularly where centre employees receive a bonus for each unemployed person who finds work). This is referred to as selection bias: accompanied persons are not drawn randomly from the unemployed population; by the same token, in the previous example, individuals who spend a lot on their health are not drawn at random from the population (they are generally less healthy to begin with).

The incidence issue

The second issue encountered in public policy evaluation is the issue of incidence: the ultimate beneficiary of the policy is not necessarily the individual targeted. This second issue is frequent in the case of taxation or subsidies/transfers. Tax incidence theory reveals that the burden of taxation may not ultimately fall upon the person who writes the cheque for it: the parties taxed may pass on the cost of the tax to others; inversely, individuals not originally targeted by a subsidy may indirectly find that they are its beneficiaries.

An illustration of this is provided by housing benefits. In 2009 this rental subsidy constituted a quarter of all benefits paid to French households. If the households officially receiving these benefits were the actual beneficiaries, housing benefit would account for over one-fifth of the reduction in inequality in living standards achieved by the welfare system.[3] However, this redistributive action is questioned by Fack (2005)[4] who considers that between 50% and 80% of these benefits in fact benefit lessors through increased rents as a result of market mechanisms (housing supply and demand): because the rental offer does not change very much in the short and medium term, demand subsidies raise rents, such that the support intended for disadvantaged tenants is in part received by the owners of rental accommodation. This example reveals that a policy of this type cannot be evaluated simply by limiting our enquiry to means or to the number of households receiving the benefit. Any analysis of housing benefit that fails to take into account its impact on rents would lead to an overly optimistic assessment (Figure 3).

3. The need to take incidence into account: the example of housing benefits



[2] Martin S., N. Rice and P. Smith (2008): "Does Health Care Spending Improve Outcomes? Evidence from English Programme Budgeting Data", *Journal of Health Economics*, no 27, pp. 826-842.

[3] Chanchole M. and G. Lalanne (2011): *Photographie du système socio-fiscal et de sa progressivité,* Rapport particulier pour le Conseil des prélèvements obligatoires.

[4] Fack G. (2005): "Pourquoi les ménages à bas revenus paient-ils des loyers de plus en plus élevés ? L'incidence des aides au logement en France (1973-2002)", *Economie et Statistique*, no 381-382.

1. The multiple effects of a policy

In the 1970s, a group of researchers organised an experiment in California, offering healthcare insurance policies to a sample of the population.[a] These insurance policies differed from one another in particular in terms of their level of personal contribution and excess payments. By monitoring households for a period of up to five years, they found that requiring a personal contribution from the insured was effective in reducing the problem of over-medication and was an effective way of limiting pointless healthcare expenditure.

However, the study revealed the possible perverse effects of such a policy, which although it certainly limited pointless expenditure, also reduced a number of instances of beneficial expenditure in the least well-off. Accordingly, by restricting themselves to the 20% of households with the lowest incomes and subjects with high blood pressure, the authors demonstrated significant variation between the blood pressure of those with full cover and those required to pay a personal contribution. In light of the impact of high blood pressure on the probability of suffering from serious cardiovascular disease in the future, this phenomenon cannot be dismissed on the pretext that the personal contribution measure was successful in its stated objective of reducing pointless healthcare expenditure.

[a] Newhouse J. (1993): *Free For All? Lessons from the RAND Health Insurance Exp*eriment, Harvard University Press, Cambridge.

Furthermore, the effect of the policy can result detrimental to those persons not benefiting from the policy. This is the case with lessees who do not receive housing benefit (they are affected by increased rents) –not to mention the additional taxation required to finance the policy.

Multiplicity of effects

The third difficulty in evaluation resides in the multiplicity of effects of a single public policy. Multiple effects may firstly arise in the same public policy field. For example, introducing a medical excess can be effective in combatting over-medication but at the same time result in a disturbing failure to access care (Box 1). In other cases, multiple effects appear in different public policy fields. For example an increase in the taxation of company profits may result in a decrease in the remuneration not of investors, but of the workforce.[5]

Where the adverse effect of a policy occurs outside the field of investigation of the evaluator, or outside the variables relevant to their study, or even their disciplinary field, there is a high risk that it will be neglected. This is why the plurality of evaluators, in terms of discipline as much as sensitivities, is primordial in order to understand all of the consequences of a public policy without being restricted to those closely linked to the initially targeted objective.

2. Public policy operating parameters

During an experiment designed to improve recourse to top-up healthcare insurance by increasing the *cheque santé* (top-up healthcare voucher), it emerged that the protocol had an impact upon the results. In particular, requested attendance at a preliminary information meeting was a discouraging factor among participants. This paradoxical and unanticipated result could have skewed the conclusions of the study, had the researchers,[a] jointly with the Primary healthcare insurance scheme in Lille, not designed their protocol with variations in the conditions and drawn lots to assign the individuals concerned to a number of groups, with and without an increase in the top-up healthcare voucher, and also with and without the initial information meeting. It appeared that requesting attendance at an information meeting could have a negative impact on the recourse to a top-up healthcare policy. The hypothesis of the authors is that although it was optional, the individuals who were unable to attend this meeting despite being invited felt that they were not justified in receiving the top-up healthcare voucher. Potentially virtuous mechanisms can also be distorted by their practical implementation.

[a] Guthmuller S., F. Jusot and J. Wittwer (2011): "Improving Take-up of Health Insurance Program: A Social Experiment in France", *Cahiers de la Chaire Santé*, no 11, Université Paris-Dauphine.

[5] See Arulampalam W., M.P. Devereux and G. Maffini (2010): "The Direct Incidence of Corporate Income Tax on Wages", *IZA Working Paper*, no 5293.

Practical implementation

Finally, the issue of practical implementation can have a strong influence over the efficiency of a policy. In certain cases, it may be tempting to reject a measure that has merely been applied under inappropriate conditions. Accordingly, the activity-based hospital tariff (T2A) probably did not yield the hoped for results in France not because of any error in its design, but rather because of an excessively detailed classification of hospital activities and an unfavourable interaction with other measures.[6] To resolve these problems, it may prove pertinent to carry out, where possible, a small-scale preliminary experiment in order to determine the operating parameters that are crucial for a public policy to be successful (see Box 2). Such an experiment in no way dispenses with the need subsequently to accurately evaluate the reform in the aggregate. In fact, prior experimentation may favour the creation of a protocol for a subsequent evaluation of policies. Specifically, there is a need to harmonise legal and administrative constraints with compliance with the economic mechanisms targeted, by leading administrations to enter into discussion with experts (and not, as is currently the case, by sequentially introducing the various stages of the decision).

Evaluation methods

Sound evaluation is in principal designed prior to implementation of a public policy, for three reasons. Firstly, there is a need for it to be based on anticipation of the expected impacts and, if possible, experimentation. Secondly, as was previously mentioned, the modes of implementation must be determined in detail. Thirdly the methodology of subsequent policy evaluation must be determined. This latter form of evaluation will be even more accurate if it is prepared in advance.

An ideal evaluation would consist in a comparison of the situation arising from the public policy with a hypothetical situation that would have arisen had the policy not come into being, with all other aspects of the socioeconomic environment being equal –a hypothetical situation termed "counterfactual".[7] However, this is impossible. The situation arising from a public policy is observable, which is not the case for the "counterfactual" situation. The difficulty in evaluation therefore resides in reconstructing what would have happened in the absence of the public policy: this situation must be constructed on an *a priori* or empirical basis, or a "counterfactual" reference group must be formed.

An ideal evaluation would consist in a comparison of the situation arising from the public policy with a hypothetical situation that would have arisen had the policy not come into being.

Random experimentation

As we have seen, a major difficulty in evaluation is linked to the fact that individuals or businesses targeted by a public policy are not taken randomly from the population: for example, they possess below average health, or are of below average employability. One way of getting around this problem is by conducting a random experiment: a group of individuals or businesses are selected by random drawings and applied a policy, whilst another group constitutes the control group. The randomised selection from a sufficiently large population ensures that the control and treatment groups are comparable: policy access is not dependent upon individual characteristics.

The experiment set out in Box 1 is a random experiment: the researchers offered free insurance to a large panel of Californian households. The participants were not free to choose the type of insurance policy offered to them. The policy, and particularly whether or not it provided full cover or required a personal contribution, was selected by random drawings.

A random experiment may prove costly, although the accuracy of the results and the budgetary savings that they may enable to be made often make them a profitable investment. The cost and complexity of random experimentation are matched by the results that they are able to yield. Accordingly, the random experiment set out in Box 1, conducted in the 1970s, serves as a reference today even though behaviours have changed since that time.

Moreover, random experiments often raise ethical issues in certain cases.[8] Certain fields are ill-suited to random experiments for reasons of fairness. For example, it would be inconceivable (and unconstitutional) to study the effect of a tax reform by subjecting various contributors at random to different taxes; others are unsuitable because they may place vulnerable subjects at risk.

Although a purely random experiment can turn out to be difficult and sometimes costly, some substitutes may also yield

[6] See Saint-Paul G. (2012): *Réflexions sur l'organisation du système de santé*, Rapport du CAE, no 103, La Documentation française and the "Commentary" of B. Dormont in the same volume.

[7] This ideal is that of medicine where a treatment is tested using two comparable groups of individuals in respect of which treatment is administered.

[8] See, on this subject, the opinion of the Ethics Committee of the CNRS (COMETS) on social experimentation: http://www.cnrs.fr/comets/IMG/pdf/07-experimentation-sociale-201001 19-2.pdf

3. Econometric methods within the context of "natural" experiments

Double difference

Because it is not possible to compare identical individuals in two different worlds (with and without the public policy), one must compare "treated" individuals before and after experimentation (we are in this case subject to situational bias) or individuals under experimentation with those not under experimentation (here we are subject to selection bias). The principle of double difference evaluation consists in a combination of the two approaches. We bring together the individuals in an experimentation group (those whose situation is deemed to have been altered by the public policy) and a control group (those whose situation has not been altered). We then compare the trends in these two groups, with the control group serving as a counterfactual reference for the experimentation group.

The increase in the tax rebate threshold for household employment which was decided upon in 2002 may be used as an example. A simple comparison of the declarations for household employment before and after this increase might lead one to believe that the measure was very effective. However, this measure was introduced when these services were in full development and the raising of the threshold occurred concomitantly with other incentivising measures (lowering of social security contributions), administrative simplifications (simplified service employment voucher) and the entrance of businesses into a market almost exclusively made up of individual workers. The taking into account of a counterfactual component (in this case those households not affected because they were previously situated below the former threshold or above the new threshold) enables the specific effect of the measure to be isolated since members of the control group are affected just as much as those of the experimentation group by the other incentivising measures. Accordingly it is found that increasing the threshold did raise the demand for home services but was only marginally responsible for the development of the sector.[a]

Matching methods may further improve evaluation. These consist in identifying similar individuals within the control and experimentation groups. The comparison of trends in the relevant variables is then no longer made between the control and the experimentation groups but rather individually between the mixed sub-sets taken from these groups.

Regression discontinuity design

Another method consists in identifying a discontinuity in the right to benefit from the measure and in carrying out the evaluation at this level only: this is the principle of regression discontinuity design. Fack and Grenet (2010) used this method to estimate willingness to pay for education, based on a discontinuity in school catchment areas.[b] The price of accommodation per square meter depends on the neighbourhood and the quality of the accommodation; it is relatively stable where location and quality are equal. Accordingly, by comparing apartments of identical quality, on either side of the same street -therefore in the same neighbourhood– but where the street addresses sent children to different schools as a result of the school catchment areas, they were able to measure the price that parents were willing to pay to send their children to one school rather than another.

Regression discontinuity design therefore consists not in comparing all of the individuals under experimentation with those who are not, but only those who are very close to the threshold that determines to which of the two groups they are assigned. Assuming that the characteristics of the individuals are continuous, the individuals very close to the threshold on one side (and therefore not experimented upon) are identical and therefore comparable to those individuals very close to the threshold on the other side (and therefore experimented upon).

Another example is the use made by Piketty and Valdenaire (2006) of thresholds for creating new classes in order to estimate the impact of class size on school performance.[c] Class sizes are not determined randomly: they are not the same in urban and rural areas, and within the same school children are not distributed randomly among the classes. It is therefore difficult to determine the impact of class size on school performance. The two authors make use of the rule according to which a French second year primary school class (CE1) may not number more than 30 pupils: when a new pupil arrives (random event) in a cohort of 30 pupils, an additional class is created, and pupils are then taught in classes of 15 or 16. This event creates a discontinuity which may be exploited to measure the impact of class size on school results.

Instrumental variables

One final method consists in finding a variable considered to be "instrumental" to delineate control and experimentation groups. This is a variable that is closely correlated with the fact of being "addressed" (by the public policy), but without any direct bearing on the variable of interest (the result of the policy) and cannot be manipulated by individuals. This method has been used to estimate the impact of maternity on participation in the labor market, which is useful for calibrating policies encouraging mothers to take up work. The problem here is that the choice of the number of children is influenced by the status of the mother on the employment market (employed, unemployed or inactive). To work round this problem, Angrist and Evans (1998)[d] separated a homogenous group of women with at least two children according to whether the first two children were of the same sex or of different sexes. In theory this "instrumental" binary (same sex, different sexes) variable has no direct influence on the participation of women in the employment market. However, women whose first two children are of the same sex have a third child more often than the others, for exogenous reasons, not owing to differing individual characteristics or their status in the labor market. The authors then observe that women whose first two children are of the same sex participate significantly less in the employment market than women with two children of different sexes which they interpreted as being the causal effect of their having a third child.

[a] Carbonnier C. (2010): "Réduction et crédit d'impôt pour l'emploi d'un salarié à domicile, conséquences incitatives et redistributives", *Économie et Statistique*, no 427-428, pp. 67-100.

[b] Fack G. and J. Grenet (2010): "When do Better Schools Raise Housing Prices? Evidence from Paris Public and Private Schools", *Journal of Public Economics*, no 94, pp. 59-77.

[c] Piketty T. and M. Valdenaire (2006): "L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français. Estimations à partir du panel primaire 1997 et du panel secondaire 1995", *Les Dossiers Enseignement Scolaire*, no 173, National Education Ministry.

[d] Angrist J. et W. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size", *The American Economic Review*, no 88, pp. 450-477.

very satisfactory results. At the point at which a public policy is decided upon, this consists in not implementing it in a single stage throughout the territory but staggering its implementation in a number of waves, for example by groups of *départements* (administrative areas). The *départements* of each wave then need to be chosen such that each is as closely comparable as possible to the others.

This type of experimentation had been envisaged for the replacement of French Minimum Income (*Revenu minimum d'insertion, RMI*) by an Earned Income Supplement (*Revenu de solidarité active, RSA*). The measure was first applied in the *departement* of Eure, and then in 25, 34 and finally 40 *départements* before being fully rolled out across the nation. However, the experiment was not conducted with sufficient consistency and evaluation was not of the calibre that had been anticipated.

Random experiments can yield very reliable results provided that they are prepared in advance, either by defining the test groups, or by sequential implementation of the policy. Where this is not the case, other evaluation methods must be envisaged.

Natural experimentation

A "natural"[9] experiment consists in comparing groups of individuals (or businesses) which are found to be unintentionally separated in terms of access to the policy under consideration. Individuals excluded from benefiting from the public policy act as a "counterfactual" against the actual beneficiaries.

The difficulty in evaluating a natural experiment is the validity of the "counterfactual" component, i.e. the comparability of the experimentation and control groups: an apparent similarity between the groups compared does not exclude the presence of bias in the evaluation. A number of econometric techniques have consequently been developed to ensure the comparability of the experimentation and control groups. Box 3 sets out some of these techniques.

The limits of random and natural experiments

The techniques set out above are relatively recent and their ability to provide explanations may be relied upon. However, their predictive abilities have been disputed on the grounds

[9] The qualifier "natural" does not necessarily designate a connection with nature but simply an unintentional character.

that behaviours depend upon a constantly changing socioeconomic environment. Accordingly, behaviours may differ between the reaction to a small-scale experiment and the reaction to implementation of the actual policy, owing to the fact that the aforesaid policy more radically alters the economic context. For example, the random experiment of the RAND Corporation (Box 1) examines the effect of insurance on the consumption of care in a small number of subjects. If we consider a public insurance policy of greater scope such as Medicare,[10] we may observe a much greater increase in spending on care: the fact that the insurance encompasses more individuals and substantially increases the financial resources of the sector leads to a supply-side care reaction, the construction of new hospitals and an increase in medical research. Methods combining estimation in natural experimentation and more general models are currently being developed in order to correct this handicap.[11]

The need for reliable data

To conduct evaluations, the availability of databases that are both exhaustive and reliable is imperative. Researchers rarely have the means to conduct surveys themselves. Fortunately, the required data already exists, mostly in various government databases. It is therefore important to set in place institutions and procedures that enable researchers to make use of this data whilst safeguarding the rights of individuals and businesses whose information is stored in these databases. Secure protocols already exist in France, such as Secure Data Access (SDA), requiring agreement for each study from the Statistics Confidentiality Commission (*Commission du secret statistique*). However, two lacunae are still acutely felt.

The first is that much data remains inaccessible, in particular health insurance and tax data. Yet, this data is essential for a number of evaluations, for several reasons. Firstly, there are many tax policies and these have been little studied due to a lack of data. Access to tax data would enable great progress to be made with these types of policies. Also, because tax data is so rich, it can be deployed in the evaluation of non-tax policies. Therefore, evaluators need to be allowed to work with these databases whilst ensuring confidentiality for taxpayers. This would take the form of safeguards such as SDA and anonymisation of the databases, whilst keeping observations codified to ensure that it is possible to make up panels. This is technically straightforward and inexpensive.

The second lacuna concerns the possibility of comparing government files or surveys. In effect, even if the information contained in the databases that are accessible to evaluators were to expand, it would not always be usable. To study the behaviour of mothers in terms of the labour supply, for example, information is needed on a woman's children, which is found in one database, and on her participation in the employment market, which is found in another. If we are unable to match the databases, then the information contained in each database becomes useless. Simple, reliable and inexpensive ways exist of matching data whilst complying with data anonymity.

How can different public policies be compared?

The evaluation of a public policy may lead to a clear-cut conclusion: the policy is inefficient as regards the stated objective, and is even counterproductive. However, often the assessment is more nuanced: the policy is effective but appears costly in relation to the results obtained. To conduct a comparison with other public policies, particularly those active in different public policy fields, the benefits must be converted into a metric that makes them comparable with both the costs and the benefits of other public policies. In practice, a monetary value must be ascribed to non-monetary benefits such as air quality, longevity or health.

On the face of it, this may appear shocking, but it is the only means of making the criteria used for public decision-making explicit. These monetary values may be defined protectively, as has often been the case in the domain of road safety. However, it is preferable to seek to identify the preferences of individuals from surveys in which they express their willingness to pay for example for an improvement in water quality.[12]

Structuring of evaluation

Public policy evaluation is not just a matter of data and technical expertise. The policies evaluated are often complex and generally effect a redistribution within society. These characteristics impose the need to organise the evaluation with a great deal of thoroughness at the various expertise levels.

[10] Finkelstein A. (2007): "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare", *The Quarterly Journal of Economics*, no 22, pp. 1-37.

[11] Attanasio O., C. Meghir and A. Santiago (2012): "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA", *Review of Economic Studies*, no 79, pp. 37-66.

[12] One very important point is the manner in which we report the willingness to pay of individuals. A number of studies, on the environment, and also on health, show that public decision-making can often be significantly altered by the weighting schemes. See Anthoff D., C. Hepburn, R.S.J. Tol (2009): "Equity Weighting and the Marginal Damage Costs of Climate Change", *Ecological Economics*, no 68, pp. 836-849; Fleurbaey M., S. Luchini, E. Schokkaert and C. Van de Voorde (2013): "Évaluation des politiques de santé: pour une prise en compte équitable des intérêts des populations, *Économie et Statistique*, Forthcoming.

Technical expertise and administrative expertise

Although technical expertise is indispensable in order to avoid the evaluation pitfalls set out above, we could not do without administrative expertise on the practical implementation of policies and the operation of the public sector or governmental institutions that manage them. Administrative expertise not only enables an identifying strategy and counterfactual scenarios to be established, but also allows us to discern, within the results, what arises from the policy's general principal and what owes itself to its practical implementationThese two types of expertise –technical and administrative– must collaborate not only during the evaluation phase itself but equally and particularly upstream of it, if possible prior to implementation of the policy to be evaluated. This must enable us to adapt certain legal mechanisms to make it possible to evaluate it more accurately. It is also during this *a priori* coordination phase that the decision may be taken, depending on the legal options, to stagger implementation of the measure to construct a pertinent counterfactual component prior to evaluation. Finally this *a priori* coordination could enable failings in implementation to be avoided that would be liable to render ineffectual a public policy that is in principal beneficial.

Independence of evaluators

No matter how thorough it may be, an evaluation remains subject to scientific uncertainty: the results are conditional upon the validity of the methods (choice of counterfactual component, generalisation of results, etc.). Yet, in order for the public policy evaluation to be useful, it is important that these results are credible: that the hypotheses are presented transparently, without leading one to suspect that some of them have been concealed. Transparency and credibility require the independence of the evaluators. The difficulty then lies in getting institutional and scientific partners to work together whilst safeguarding the independence of the evaluation. Obvious conflicts of interest exist where the evaluation is conducted by administrations, ministries, directorates or public institutions tasked with designing or applying a public policy. The same institution cannot be both judge and judged. This is not the only independence issue, however, and care must be taken not to create *de facto* dependency during the process of appointment of evaluators, or by blocking the publication of results.

The evaluation timeframe is not the same as the policy timeframe. This divergence in the timeframe is brought about by a number of factors. Firstly, most evaluation methods require longitudinal data. We therefore need to wait until sufficient data has been compiled in order to ensure the validity of the evaluation. Moreover, the evaluation itself takes time, from the appointment of evaluators to the discussion of results, and also the choice of methodologies and the statistical work. In this field, haste is often the enemy of precision and exhaustiveness. In particular, hypotheses must be subjected to in-depth peer review. This an *ex post* validation, which is the truly scientific method, must not be neglected in the case of public policy evaluation.[13] Rushing an evaluation results in a reduction in its credibility. This makes public decision-making more difficult and potentially longer.

Dissemination of results

The freedom to disseminate results is a key condition for the independence of the evaluators. Specifically, access to data must not be subject to monitoring rights by the administration providing the data. Any practice that runs counter to this freedom to disseminate, by placing pressure on the evaluator in terms of their results, would contravene the independence requirement for this latter, in addition to the fact that it would exclude any scientific debate on the method and results.

> The freedom to disseminate results is a key condition for the independence of the evaluators.

The dissemination of results must be accompanied by a cross-comparison with other evaluations emerging, where applicable, from other disciplinary fields (see below). This cross-comparison must take the form both of the publication of the different results and their critiques, and also the organisation of debates, and even consensus conferences. This is to enable citizens to be better informed and to better understand the various effects of a public policy. To this end, methods for hierarchising the robustness of the results may be brought into play (scientifically established proof, scientific assumption, low level of proof).

Plurality of evaluators and interdisciplinarity

Since a single public policy often has multiple effects, it is necessary to have at our disposal a number of evaluations corresponding to different approaches, disciplines or sensitivities. The French Competitiveness and Employment Tax Credit (*Crédit d'impôt compétitivité emploi*, CICE, Law no 2012-1510 of 29 December 2012) serves as an example in this regard. When evaluating this measure, a number of relevant variables may be envisaged. Firstly, an impact study on the commercial position of France in relation to foreign countries would seem

[13] This peer discussion phase safeguards the independence of the choice of method; it also facilitates a clear separation between policy evaluation and decision-making. See the opinion of the CNRS Ethics Committee, *op. cit*.

4. Examples of organisation of evaluation abroad

Various bodies have been established to coordinate the evaluation of public policies abroad. One illuminating comparison is that of the Institute for Fiscal Studies (IFS) in the United Kingdom and the Government Accountability Office (GAO) in the United States.[a] The GAO is not independent in nature, and was not initially involved with evaluation. It reports directly to the federal government and its role when created in 1921 was to audit the finances of government agencies. Auditing is very different from public policy evaluation but over time, the remit of this institution was broadened. A number of researchers in the social sciences, and joint work undertaken with academics, have been incorporated into the primarily juridical remit of the GAO. Its objective is to inform Congress and citizens about government policies, to enable Congress to perform its legislative role better, and be able to oppose, where needed, the executive in an informed way. It does so namely by monitoring ministerial evaluations in accordance with scientific criteria (validity of counterfactual components) and institutional criteria (separation of commissioner and evaluator, independence of this latter, automatic publication of results). In order to ensure its independence, the GAO is managed by the "United States Comptroller General" whose mandate is long (15 years), and which cannot be shortened and is non-renewable.

Transposition to France would not be easy, however, and similar independance may be difficult to obtain. The United Kingdom, which, like France, has a strong executive based on a powerful administration, has set in place a different system altogether. Accordingly, the IFS, which has the status of a non-governmental association, is independent by its very nature. In order to prevent dependency upon established interests, financing is drawn from numerous subsidies from institutions and businesses. The only core subsidy is from the Economic and Social Research Council (ESRC), the public agency tasked with financing social sciences research in the country. Scientific competency is safeguarded by the recruitment of social sciences researchers and long-term collaboration with academics. Its roles are principally centred on public policy evaluation and the explanation of measures whose complexity jeopardises transparency. The results serve to advise members of Parliament, the Government, and various groups from civil society. Finally, the IFS has assigned itself a key role in respect of the general public, with publications for information purposes in the press.

Much may also be learned from the case of Australia. Lamenting the fact that control of expenditure should take priority over the evaluation of performance, the Australian government determined to foster in its ministries a true evaluation culture at the end of the 1980s. Each minister was required to submit to the Ministry of Finance an annual evaluation plan, enabling all of its policies to be evaluated every three to five years. The results were then made public. The Finance Minister, and in particular the Australian National Audit Office (ANAO) monitored these evaluation practices. Specifically, the ANAO acted both as an evaluation consultancy body and itself evaluated the quality of these evaluations. The result was effective evaluation of policies and significant factoring in of the results in new policy proposals. However, ANAO stated, in its 1997 report[b], that communication around the methods and results of the evaluations conducted by ministers in charge of policies was insufficient.

---

[a] For an in-depth discussion, see Ferracci M. and E. Wasmer (2011): *État moderne, État efficace* (Modern State, Efficient State), Odile Jacob.
[b] ANAO (Australian National Audit Office) (1997-1 998): "Program Evaluation in the Australian Public Service", *AGPS Performance Audit Report*, no 3.

to be an obvious requirement, as well as an impact study on employment in quantitative terms. However, other evaluators might take an interest in other issues, such as the impact of the policy on the structure of qualifications, of careers within industry or on employment conditions. We might also evaluate the impact on the financing of social security from a more general standpoint, and on its acceptability.

This example illustrates the importance of pluralism not only in the *a posteriori* evaluation, but also in its *a priori* preparation. The upstream coordination phase must enable relevant variables to be defined, changes in which we wish to measure in view of the policy under study, and enable each of the consequences of the measure to be evaluated. It is therefore important that this *ex ante* coordination be pluralist in terms of methods, disciplines and sensitivities. This *ex ante* pluralism is more complex to set in place than *ex post* pluralism, namely because this latter may be obtained through the juxtaposition of evaluations issuing from different expert groups. The initial coordination phase being unique, even more sus-

tained attention must be given from this stage onwards to ensuring the existence of pluralism.

The evaluation triptych

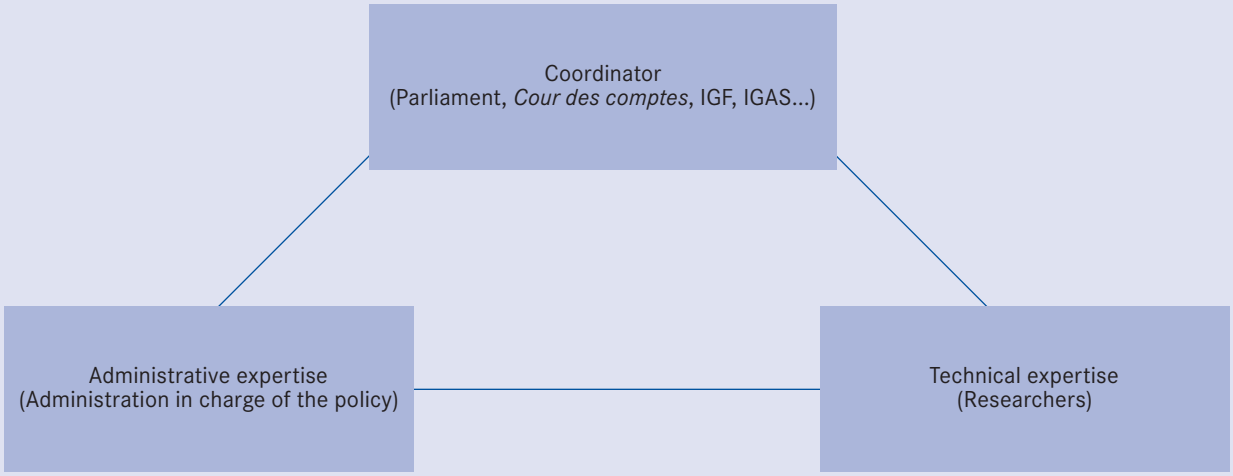
Contrary to other developed countries (see Box 4), France has little experience in public policy evaluation within the meaning defined in this Note. Sound evaluation should be based on the triptych form of coordinator, the administrations concerned and independent experts:

- *coordination* of the evaluation should be provided by an institution that is external to the executive. Democratic logic dictates that it should be Parliament that is tasked with commissioning these evaluations. This would mean vesting it with the technical ability, i.e. the personnel, to genuinely coordinate implementation of evaluation, its appraisal and its dissemination to members of parliament and the general public. The *Cour des comptes* is another candidate, with sufficient institutional weight to commission independent evaluations.[14] Whatever the choice of institutional commissioner, it must coordinate preparation of the evaluation, ensure compliance with a plurality of approaches, and check that all measures likely to facilitate evaluations have been taken (particularly data access). It must also organise discussion and the dissemination of results. Evaluatory bodies should be selected on the basis of public tenders;

Sound evaluation should be based on the triptych form of coordinator, the administrations concerned and independent experts.

- *the administrations concerned* should provide their institutional expertise. Collaboration with evaluators must be without pressure, under the oversight of the commissioning bodies. Ministerial statistical departments possess technical skills in the area of evaluation, in addition to institutional skills. They can therefore organise parallel evaluations to the independent evaluations and participate usefully in discussions on the results. However, they must facilitate execution of the independent evaluations, particularly by providing full and informed access to the data;
- *experts* must provide their scientific skills as evaluators. Their independence must be safeguarded amongst other measures, by rotation, so as to prevent any skewing based on the previous results of evaluations. Experts must be bound by statistical confidentiality constraints and be transparent as regards any

4. The evaluation triptych



[14] The coordinator may also come from the administration as long as it is not the administration in charge of the policy under evaluation. The General Inspection of Finances and the General Inspection of Social Affairs come to mind.

ancillary activities that may cause conflicts of interest to arise. It is vitally important that they work together with other disciplines, particularly during the preparatory phases and those of results dissemination.

Conclusion

Although requiring the combination of technical expertise, administrative expertise and thorough organisation to ensure independence and pluralism, public policy evaluation is nevertheless not beyond the capabilities of a government that is determined to sift through its public policies. Three *sine qua non* conditions for the evaluation to be successful and credible must however be underlined: data access, expertise time and the publication of results. These conditions must not be viewed as constraints, but rather as the key ingredients of a credible evaluation, upon which the decision-making process can genuinely be based with the utmost transparency.