Société Française de l'Évaluation

How can we improve the usefulness and effectiveness of evaluations?

The role of quality standards

The two main objectives of evaluation are common knowledge: on the one hand, report to taxpayers and citizens on public actions, which makes it possible to ensure "accountability" and, on the other hand, learn from the past in order to improve future actions. Improving the usefulness and effectiveness of evaluations therefore helps to improve the effectiveness of a public policy. In light of these considerations, improving the usefulness and effectiveness of evaluations improves the effectiveness of public actions and policies.

With this issue in mind, national evaluation societies, networks of evaluators and international standards organisations, first and foremost of which is the OECD Development Assistance Committee for evaluations of development actions, have sought the best method for achieving this result by defining quality standards or publishing guides to best practices.

Laying down these standards was the result of a long process that started in the sixties, in particular in the field of education, through the "Standards for Educational and Psychological Tests and Manuals; American Psychological Association". This process was then developed in the seventies until it was possible to define standards in a particularly detailed, rigorous manner, in a similar fashion to the standards that the Swiss evaluation society developed a few years later.

All in all, we can consider that the effectiveness of an evaluation essentially depends on four founding principles:

All evaluations must be useful; in this regard, there is no point in evaluating everything. Performing an evaluation is costly and time-consuming. For example, spending EUR 30,000 to evaluate a project that only cost EUR 20,000 is clearly absurd. We therefore need to avoid the syndrome of wanting to over-achieve. Secondly, the singularity of an evaluation is based on the time required to perform it properly. This time, which is long, is in contrast to the short time needed for a public or political decision. It is therefore essential for the evaluation to remain connected to the short timeframe of politics, by linking the performance of evaluations to the decision-making process or circuit. For example, carrying out an evaluation upstream of a decision-making process is often relevant; performing an evaluation with no timeframe orientation runs the risk of suffering from this time lag and ultimately being of no use.

The second founding principle is that an evaluation must be feasible. The aims of the evaluation must be precise, targeted and have a limited scope. The party that commissions an evaluation can tend towards excess, i.e. to want to evaluate as many things as possible going as far back in time as possible. Experience shows that an evaluation must be targeted in time and space, with the aim of defining measurable, limited objectives. In this regard, policy evaluations are more complex, insofar as they often have a significant longitudinal aspect. Identifying the objective, the principal and the implementation makes it possible to ensure the proper feasibility of an evaluation.

65 avenue des Gobelins – 75013 PARIS
Tel.: +33 1 45 41 58 40
Website: www.sfe.asso.fr E-mail:evaluation.sfe@wanadoo.fr

Thirdly, an evaluation must be run in compliance with ethical rules and codes of conduct. Of course, it is obvious that an evaluator must not have been directly or indirectly involved in the management of a project, programme or policy. However, the evaluation team must also follow a set of rules that makes it possible to ensure its credibility and ultimately its professionalism. In this regard, the accuracy of assessments and information bear witness to the validity and precision of the data and the analysis. They are the final founding principle. Observations must be based on data and facts that are checked and crosschecked and which, once processed, enable objective analysis and therefore objective judgment. This then makes it possible to formulate impartial findings.

In this context, in the field of evaluating development activities, the OECD Development Assistance Committee quality evaluation standards were adopted in 2006 and sum up these principles in ten main criteria:

- *rationale, purpose and objectives of the evaluation*
- *evaluation scope*
- *context*
- *evaluation methodology*
- *information sources*
- *independence*
- *evaluation ethics*
- *quality assurance*
- *relevance of the evaluation results*
- *completeness*

Following the application of these standards in most of the Member States, the DAC Working Party on Evaluation looked to draw conclusions from this initial application of standards. This exercise was started in 2009 with the aim of revising the standards, where applicable.

There are two separate objectives in implementing such standards:

On the one hand, the existence of these standards and their formalised nature allows operations departments to have a modus operandi *that develops their knowledge and know-how in the area of evaluation. Evaluation is not a random, subjective process that aims to perform some kind of disguised audit or inspection. Drafting these standards therefore enables the main questions to be answered, by providing concrete, objective, substantiated and therefore professional answers.*

On the other hand, one of the weak points of evaluation effectiveness is measuring the quality of consultants' work. The operations departments, i.e. the departments that manage a project, a programme, an instrument or a policy may be tempted to cast doubts on the credibility of a study if the study highlights the ineffectiveness of an evaluated action; conversely, an objectively poor-quality evaluation (lack of methodological rigour, data reliability not checked, autonomy leading to a battle of wills, etc.) cannot be sanctioned. Lastly, both parties (the manager and the evaluator) may be frustrated insofar as their interaction is limited to the delivery of the report, with no assessment of the usefulness of their work together or the efforts made.
In order to answer all these questions, on the basis of these standards, an assessment grid for the consultants' work has been drawn up. This grid is based on four or five criteria, for

which a mark is given[1] (from very good to very poor) and makes it possible to assess the work performance of the evaluators for each criterion.
Marking must be performed collegially. Having only the party that commissioned the evaluation perform the marking should clearly be avoided, as this introduces a power struggle (the evaluated party evaluates the evaluator); in the same way, we can consider that the evaluated unit or department directly marking the consulting firm is not ideal, insofar as it places the evaluated structure in a position of arbitration. The most satisfactory solution seems to be to have a steering committee, provided that its membership is diversified and ensures that the evaluation is independent. Under these conditions, each steering committee member provides a personal assessment that is as impartial as possible of the quality of the work provided. Using this grid makes it possible to reach three objectives, for which the results are particularly interesting.

Firstly, its initial function is to enable the consultants' work to be assessed. In doing this, the consultants are aware of the steering committee members' opinion and can therefore obtain feedback on the work performed. This method constitutes a valuation of the efforts made and allows the consultants themselves to learn from their performance and improve their work in the future.
Secondly, collegiality is assurance against bad faith or using the results of an evaluation as a tool, in a positive or negative way. The management department cannot use the evaluation as a tool if the findings are favourable solely due to the complacency of the consultants; conversely, if the findings of a study are critical vis-à-vis a management department, this department cannot exempt itself merely by attempting to discredit the quality of the evaluation. Under these conditions, collegial marking makes it possible to highlight focal points, i.e. any divisive issues concerning the evaluated action. In so doing, this grid makes it possible to provide an objective approach to a subjective perception of the quality of an evaluation. The collegiality of assessments is particularly important here. The involvement of representatives of other ministerial departments, representatives of civil society, the private sector and academic or institutional experts makes it possible to optimise the independence and impartiality of judgements.
Lastly, drawing up such a grid is extremely useful for evaluation units or departments. Without such a grid, there is scant legal justification for lists of "good firms" and "bad firms", due to a lack of objective, transparent marking criteria. This means assessments are oral, necessarily more subjective and based on people's memories, which by definition are fallible, in particular if there is a high level of staff turnover. It is impossible for this method to be satisfactory. The use of public, published marking grids therefore makes it possible to assess the quality of consulting firms' work over time and to provide an objective, collegial, transparent basis for performance.

Evaluation is a learning process. The usefulness and effectiveness of public actions or policies result from evaluations with improved formalisation and evaluations that are more professional. Using quality standards seems to provide encouraging results in terms of the formalisation of evaluations.

However, there are still numerous challenges. Quality standards have not yet been applied or implemented in full by all countries, all ministries or all agencies. Yet, complete application is a sine qua non *for achieving the expected results. For example, the absence of a diversified, competent steering committee does not make it possible to ensure the independence and therefore the relevance of its judgments. There is not just one method for achieving these objectives, as socio-cultural context must be taken into account, and any variations in methods must not prevent them from remaining cohesive as a whole.*

[1] Two methods are possible: using a marking scheme with an odd scale of 5 (--/-/=/+/++), which aims to be as precise as possible, or a marking scheme with an even scale of 4 (--/-/+/++), which eliminates the average mark and makes it possible to distinguish between a reasonably good performance and a reasonably bad performance in terms of reaching objectives.

Author: Benoît Chervalier

Bibliography: Benoît Chervalier is the head of the evaluation unit for the development activities of the Treasury and Economy Policy General Directorate (Ministry for the Economy, Finance and Employment) and Vice Chair of the OECD Development Assistance Committee Evaluation Network.