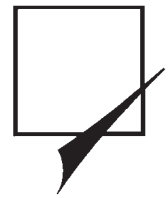
Evaluation

SAGE Publications (London,
Thousand Oaks and New Delhi)
DOI: 10.1177/1356389004043135
Vol 10(1): 58–76



Theory-based Evaluation and Types of Complexity

NICOLETTA STAME
University of Rome 'La Sapienza', Italy

Theory-based evaluations have helped open the 'black box' of programmes. An account is offered of the evolution of this persuasion, through the works of Chen and Rossi, Weiss, and Pawson and Tilley. In the same way as the 'theory of change' approach to evaluation has tackled the complexity of integrated and comprehensive programmes at the community level, it is suggested that a theory-oriented approach based on the practice of realistic cumulation be developed for dealing with the vertical complexity of multi-level governance.

KEYWORDS: complexity; institutional evaluation hierarchies; integrated programmes; multi-level governance; theory

Black Boxes and Evaluation Deficits

Since its inception with the 'War on Poverty' programmes in the US, evaluation has been plagued with the 'black box' problem. The black box is the space between the actual input and the expected output of a programme. Moved by the need to tackle serious social problems, programme designers often gloss over what is expected to happen, the how and why, when an input is put in place; and evaluations do the same concentrating on measuring outputs, whilst attributing the observed difference to the input. All this is hardly informative for a policy design wishing to build upon previous experience. However, for a long time evaluations have coexisted with black box programmes, and have tried to cope with similar shortcomings by developing sophisticated methods for measuring the distance between objectives and results.

Nowadays the evaluation community has become more and more concerned with the challenge of how to understand 'what works better for whom in what circumstances, and why' (Pawson and Tilley, 1997) to improve policy decisions and public sector practice. Theory-oriented approaches reproach the previous, method-oriented approaches for being ineffective, given their inability (or unwillingness) to 'open the black box'. In recent times, theory-oriented approaches have flourished, most notably of the 'theories of change' kind (Kubisch et al., 1995), with respect to the evaluation of complex programmes.

And while the black box of programmes is being opened, a further missing dimension is moving centre-stage: the 'evaluation deficit' at state and EU levels.[1] The evaluation deficit refers to the unsatisfactory situation in which most evaluations, conducted at local and other sub-national levels, provide the kind of information (on output) that does not immediately inform an analysis of effects and impacts at higher levels, i.e. whether global objectives have been met. Or conversely, impact assessment is not corroborated by an understanding of the working of programmes. Consequently, evaluations are under-utilized or not utilized at all. Although this has not been diagnosed as a black box problem, the vertical predicament shares many points in common: generalizations are not easily drawn and it is not clear how lessons learned at the local level can be useful at the higher level. There is an even more striking similarity: the issue keeps being tackled as if it were either a matter of a lack of government capacity or of bad quality evaluations in terms of methods.

The aim of this article is to explore whether it would be possible to deal with the evaluation deficit at the EU and state level by utilizing the experience of theory-based evaluation at the programme level. The article is divided into two parts. The first part examines the way theory-oriented approaches have tackled the black box problem at the programme level. The second part highlights the problem of the evaluation deficit and tries to understand how current ideas stemming from the theory-oriented movement can contribute solutions. This part will draw also on personal experience in evaluating the European Community Initiative 'Urban'.

Method v. Theory

Evaluation was born when the first programmes of the US War on Poverty were launched. Yet, its relationship to programmes is still far from being settled. Programmes encompass actions with desired outcomes. Although the latter are shaped by values, programmes present themselves as the best way to reach those outcomes, hence as 'rational' actions, unfolding along a chain of 'objectives – means (inputs) – outcomes'. At the same time, policy makers do not adopt programmes solely because of theories tested by social researchers. They also follow their own bias, opinion-polls, assumptions, and the fad of the moment, which it is taken for granted will work. Of course, the situation is not always so bad. The accumulation of knowledge about the working of programmes has increased the chances of good programmes being repeated and of bad ones being improved: the spread of 'integrated' programmes, and of 'flexible' procedures is witness to this.

The 'original sin' of mainstream evaluation (that with a positivist imprint) lies in choosing to play a low-key role. Neither wanting to enter into the 'value' problem (thanks to a value-free stance), not wanting to discuss the theoretical implications of programmes, evaluators have concentrated their efforts on developing a methodology for verifying the internal validity (causality) and external validity (generalization) of programmes.[2] No wonder so much energy has been spent not only in developing evaluative methods in general (the pride of the profession), but methods suited to test programmes framed in that way,

i.e. randomized experimental designs[3] that, according to Chen and Rossi (1989: 301) 'fail to formally or explicitly specify theory'.[4]

There were several consequences of this low-key inception of programme evaluation.

1. Not discussing programme theories amounted to warranting programmes with 'absolute rationality' (assuming that needs are known, decision makers are informed about the risks and opportunities implied in each option, decisions are taken with the aim of maximizing the gains from existing resources) at a time when most policy analysis had accepted bounded rationality and incrementalism as its new paradigm (Etzioni, 2001).[5]
2. If programmes could be regarded as 'rational actions', then politics was seen as a disturbance or interference and the political context itself never became an object of inquiry.
3. Concentrating on verifying the validity of programmes whose theoretical implications were not questioned, led evaluators to believe that the outcome of evaluation would be 'instrumental' use: saying that something worked or did not work. This provided the commissioner with a clear answer about 'go/no go' and the decision maker would then follow suit.

Why did evaluators dismiss their responsibility as social scientists even in a country where the pragmatist tradition of Dewey was still widespread?[6] Homage to a badly understood idea of practice? A too strict separation from other disciplines (like policy analysis, organization theory, etc.) thanks to an ill-conceived idea of disciplinary specialization?

Indeed these issues were discussed even before the emergence of theory-oriented approaches.[7] 'Responsive evaluation' (Stake, 1980), 'fourth generation' evaluation (Guba and Lincoln, 1989), and the 'reformist' approach of Cronbach et al. (1980), have all in their own way addressed the search for generalization, and have advocated that evaluations should consider the context in which programmes are enacted, and the different interests and views of stakeholders. For his part, Scriven proposed that evaluations should be 'goal-free', and his 'logic of evaluation' reintroduced values, in a pragmatic fashion.[8] It is true however, that, for different reasons, none of them thought programme theory crucial: they have mainly fought against positivist approaches on methodological grounds.[9]

And here enters theory. Theory-oriented evaluations present themselves as a new wave vis-à-vis method-oriented evaluations. In this new wave, what changes is the attitude toward methods. There are no more paradigm wars that are immobilizing the field; nor contention about pre-planned multi-method evaluations. All methods can have merit when one puts the theories that can explain a programme at the centre of the evaluation design. No method is seen as the 'gold standard'. Theories should be made explicit, and the evaluation steps should be built around them: by elaborating on assumptions; revealing causal chains; and engaging all concerned parties in this exercise. It is expected that theory-oriented evaluations will help build capacities in the public sector, and educate the public to have a better understanding – and mastery – of the political process in which programmes unfold.

No Theory, Many Theories, and Actors

Theory-oriented approaches, in their turn, have taken different paths, and have addressed the black box problem from different angles.

Chen and Rossi: theory-driven evaluation Chen and Rossi, who first raised the flag of theory-driven evaluations in 1980, have proposed a kind of manifesto (1989), whose main tenet is that black box programmes are such because they have 'no theory', goals are unclear, and measures are false, with the result that evaluations are 'at best social accounting studies that enumerate clients, describe programs and sometimes count outcomes' (1989: 299). The black box is an *empty box*. And theory-driven evaluations should provide such programmes with, among other things, good social science theory. The manifesto goes on providing steps to be followed in an evaluation: studying treatment; discussing stakeholders' and evaluators' views on outcomes; examining why and how a programme fares as it does (following both normative and causal theories).

The thrust is more to provide a programme's missing theory than to discuss the way programmes exist in the world of politics. In this way, it seems that the absence of theory is seen to account for the usual way politics works; and that the deficit of politics could be corrected by a strong dose of social science theory. In this way, evaluation could make positive suggestions to policy makers and programmes would then work better.

Carol Weiss: theory-based evaluation (TBE) Weiss has addressed the problem in a different way. Programmes are necessarily confused because of the way decision making takes place; hence, insisting on aiming for 'instrumental' use is pointless. Rather, evaluators should come to terms with the way decisions are made, and try to influence decision making in a more indirect way. Good programme theory is not theory that is unaffected by politics, but a good theory of how politics works in practice. Hence, Weiss's warning: 'evaluation is a rational activity in a political environment' (Weiss, 1987). Weiss has pursued this insight in many different, and path breaking, ways: her interest in the multiple relationships between research and politics; her attention to evaluation utilization, and her understanding of cognitive use as something that is eventually conducive to better programmes.[10] It is only recently however that she has taken these arguments so far as to compare Campbell unfavourably to Lindblom and March – a stance that could be considered anathema by most evaluators. She has written (Weiss, 2000a) that Lindblom and March, with their theories of 'muddling through' and incrementalism, did more for evaluation than Campbell, the father of all methodologists.

Contrary to Chen and Rossi, then, for Weiss programmes do have theories, although as confused as a level of muddling through permits. The black box is full of *many theories*. She calls them 'theories of change' (Weiss, 1995): they take the form of assumptions, tacit understandings, etc.: often more than one for the same programme. In a remarkable effort (Weiss, 1995: 74), she lists seven assumptions, and many more sub-assumptions, that could underlie a programme on the provision of services in comprehensive community initiatives. There are

also the many ideas that implementers, stakeholders and concerned parties have about how the programme should work. All these 'theories' have to be brought to light in order to reach a consensus on which deserve to be tested (Weiss, 2000b).

In this way, Weiss suggests how to make programme theories the kernel of evaluation. Theories of change have two components: 'implementation theory', which forecasts in a descriptive way the steps to be taken in the implementation of the programme; and 'programmatic theory', based on the mechanisms that make things happen. 'The mechanism of change is not the program activities per se, but the response that the activities generate' (1997: 46). TBE should make these mechanisms clear, break down a programme in its subsequent mechanisms, and use data of different kinds to test them. In that way, TBE helps generate new theories.

Pawson and Tilley: realistic evaluation A further attack on method-oriented evaluation comes from Pawson and Tilley's 'realistic evaluation'. The characteristic of this approach is to stress what the components of a good programme theory should be: context (C) and mechanism (M), which account for outcome (O). Evaluation is based on the CMO configuration. Programmes are seen as the opportunities that an agent, situated inside structures and organizations, can choose to take, and the outcomes will depend on how the mechanism that is supposed to be at work will be enacted in a given context. Mechanisms are not infinite in number, and programmes can be grouped in relation to the mechanisms around which they are built (naming and shaming, incentives, etc.). Nor are contexts limitless, but vary according to certain characteristics (density, marginality, etc.). Contexts and mechanisms are part of middle-range theories.

> . . . the basic idea of middle-range theory is that these propositions do not have to be developed de novo on the basis of local wisdom in each investigation. Rather they are likely to have a common thread running through them traceable to [a] more abstract analytic frameworks [. . .]. (Pawson and Tilley, 1997: 123–4)[11]

According to Pawson and Tilley, the problem of the 'black box', and of the failure of experimental designs of evaluation in opening it, lies in the 'successionist' theory of causality on which experiments are based. According to this, we cannot know why something changes, but only that something has changed from status 'a' (without stimulus, without programme) to status 'b' (with stimulus, with programme) in a given case. And that is why it is so difficult to say whether the change can be attributed to the programme. The realist approach is based on a 'generative' theory of causality: it is not programmes that make things change, it is people, embedded in their context who, when exposed to programmes, do something to activate given mechanisms, and change (Pawson and Tilley, 1997: 32–4). So the mystery of the black box is unveiled: *people inhabit it*. This makes for a completely different design of evaluation. The evaluator elaborates how the mechanism could work in a given context and asks the people who could know about it to provide evidence. Different views are not obtained through consensus (as for Weiss) but through 'adjudication', a judicial metaphor for establishing what may be more worthy.

Summing up, there are differences and similarities among theory-oriented approaches. Among their similarities, theory-oriented approaches:

- base the evaluation on an account of what may happen, as understood by actors and/or interpreted by evaluators: values are accounted for in the way they help frame the actors' views, and are not ignored;
- consider programmes in their context, which includes actors' environments (embeddedness) and public service culture and behaviour;
- utilize all methods that might be suitable, without privileging any one of them, and without depending on them;
- are clearly committed to internal validity (they indeed look for causality), but nonetheless allow for comparisons across different situations.

The differences refer to the role attributed to theory. For Chen and Rossi good theories should substitute for no theory; for Weiss, better theories should substitute bad ones; for Pawson and Tilley, theories become good thanks to what actors do about them. Another difference refers to the role of context: in realist evaluation this means the sociological characteristics of an environment; the other authors conceive it as contingent.

Complexity and Multi-level Governance

Theory-oriented evaluations have helped open the black box of programmes and break the resistance of method-oriented evaluations. The latter were often requested by commissioners who did not want the programme's assumptions to be questioned, and were performed by evaluators who had found their niche in methodological expertise. Theory-oriented evaluations began to be used with 'simple' programmes, where they helped clarify that either programme theories were 'too simple' (when the black box hid a mono-causal theory that proved to be wrong), or that there was something good in programmes that had yet to become apparent Indeed this happened often, because people have ideas about how to turn programmes to their own advantage, in their own context.

Recently theory-oriented evaluations have flowered under new conditions: where 'complex' programmes were being implemented, and commissioners were aware of the failure of so called gold standard methods. Weiss (1995) argues that when it is not possible to do experiments, TBE can be sufficiently compelling. The Aspen Institute Roundtable on Comprehensive Community Initiatives (Kubisch et al., 1995) has been able to offer a new platform for theory-based evaluations, in a version of 'theories of change' evaluation that follows on from the work of Weiss.[12]

But what exactly is 'complexity', and how can theory-oriented evaluation cope with it? In a sense, it is reality that is complex because:

- it is stratified, and actors are embedded in their own contexts; and
- each aspect that may be examined and dealt with by a programme is multi-faceted.

Therefore it is always difficult to say whether a single input (an incentive, a

service) caused a given output: an input never works alone. This is however, rarely considered in programmes that often ignore the simplest propositions of social science, and treat each programme aspect separately. Take the relationship between the individual, the family and the community.[13] Programmes address one or other aspect, but rarely take into consideration that individuals, families and communities may be in relationships of conflict, that there is a dynamic between them.[14]

Now we turn to the main issue: the complexity of programmes. Since policy makers have realized that reality is complex, 'integrated' programmes have been designed – in which people are expected to act in partnerships and networks – that tackle various aspects of a problem simultaneously. Typically, a community programme has economic, social, health and education components. Complex programmes aim at creating 'synergies' that will produce results that single initiatives would not have brought about in isolation. This is the philosophy behind most of the Comprehensive Community Initiatives in the US,[15] of Communities That Care and joined-up government in the UK (Sullivan et al., 2002), or the law for integrated social services in Italy (law 382/00). There are many implementers (public, private, third sector), different beneficiaries, and different activities of the same beneficiaries, brought together in 'partnerships'.

There is yet another kind of complexity, given that integrated programmes are implemented within a system of multi-level governance. The European Structural Funds, or similar European programmes are an example of this. This system is consistent with ideas of the New Public Management, especially 'principal/agent theory'. According to this, the 'implementer-agent' pursuing his/her own interests acts in such a way as to do what the 'government-principal' aimed at, but had not the information to plan, nor the competence to implement. National governments, or the EU, encompass diverse constituencies – even the most detailed planning could not take account of local specificities, and preferences. So at the higher levels (EU and national governments) goals are set, objectives are articulated (for which indicators are sought) and public resources allocated; but means or ways of attaining results are not anticipate. Recently Romano Prodi, the President of the EC, said: 'the Commission states goals, it does not state means', implying that means are up to nation states, thanks to decentralization. It is local authorities that in their turn decide upon operational programmes, distributing resources to given sectors of intervention, and choosing actions to be implemented among those for which projects had been submitted.

This multi-level system means governments can hand over the delivery of many services to non-governmental organizations (NGOs) or other third-sector groups. It may allow good local projects to come forward, but can also result in project implementers at the local level proposing a kind of activity they are already competent at, or are used to, but that is not necessarily good for the overall programme purpose. And it is up to the decentralized planning system (and to evaluation) to show whether those activities fit inside a given goal-axis, i.e. whether they have a bearing on general goals, or are just undertaken in the interests of the implementer.

Taken together, one could say that integrated programmes deal with complexity at a given level of the hierarchy, while multi-level governance deals with the hierarchy itself. If what mattered in integrated programmes was 'synergy', what matters in vertical complexity is 'subsidiarity', meaning that the higher level does not do what the lower levels can do. In the former case, evaluation will assess what characteristics of what partnerships have worked better, and why. In the latter case, evaluation will assess the roles of the different levels.[16] Subsidiarity, after all, is not a steady state, but a mutual relationship, that needs to be actively implemented if one is to elicit the latent potentialities of beneficiaries and programme operators, and thus contribute to global goals.[17]

The problem with multi-level governance is that while this system has been conceived of as a 'rational' device (it is more convenient to leave the details to the lower level implementers), it admits an even larger black box than the old programmes. Now goals can be obtained in many different ways, and there may be as many ways of filling the black box. Two main problems become relevant here.

- How can changes that occurred at the global level (i.e. in the score of indicators of objectives) be attributed to the effectiveness of any programme, if there were so many different programmes, and if programmes were integrated?
- How can outcomes be attributed to partners at a given layer, if so many actors were involved in an action?

These are typical problems of micro, meso and macro relationships, for which explanatory theories have to be worked out.[18]

Multi-level Governance and Multiple Hierarchies

Contrary to what happens with integrated programmes, in multi-level governance programme planners and evaluation commissioners have not yet become familiar with theory-oriented evaluation approaches. Vertical links have been variously taken into account, for example through:

- multi-layered institutional arrangements for commissioning evaluations; and
- evaluation guidelines proposing ways of dealing with causality.

My contention is that the former do not help the circulation of knowledge needed for assessing success along the vertical dimension, and the latter are unable to deal with this kind of complexity.

To take the institutional arrangements first. EU Structural Funds are evaluated through a vertical, multi-layered evaluation architecture (EU, 2000). At the EU level, ex post evaluations are contracted out to independent evaluators who, presumably, will be asked to assess what changes took place in score of indicators of global objectives (global impacts). At the state level, evaluations are contracted out to independent evaluators and whose main focus is on attainment of intermediate impacts, and performance on process indicators. At the local

level, nothing is officially stated, but it is assumed that the implementation of single projects is monitored, and self-evaluation will assess outputs. Each level is self-contained; evaluations take place at different times, and are performed by different professionals with different methods. There is a risk that the findings at each level are not communicated upwards or downwards, such that it may be hard to attribute outcomes or impacts to any one actor or mechanism.

The aim of the best known guidelines for the evaluation of the EU Structural Funds, the MEANS guides (vol. 1: 93–5) is to take the complexity of multi-level governance into account. There is an attempt to understand causal links, through the combination of the 'hierarchy of objectives' and of the 'logical diagram of expected impacts' (MEANS, 1999, vol. 1: 93–5), and a search for generalizability, via 'best practices'.

The 'hierarchy of objectives' works top-down, and de-composes global objectives at the higher level into specific objectives of lower levels, cascade-like. The 'logical diagram of expected impacts' works bottom-up and re-composes the results obtained at the lower level of interventions to the higher level of global impacts.[19] Evaluations are expected to find out how single projects implemented at the local level fit into this 'ladder'. What is proposed is the articulation of the official programme theory into a series of virtuous linear chains linking the results of a local intervention, its effects on the performance of a national programme, and the latter's contribution to the impact of EU policies. 'Best practices' in their turn are ways of acting at a local level that are proposed as models for generalizations in a further virtuous and linear chain.

How do these approaches address the problem of 'establishing goals not means'? It can be argued that they allow for two alternatives: either the evaluator is only concerned to measure global impacts, regardless how they are attained, or s/he has to find out how results can be attributed to different means.

If evaluators concentrate on objectives, they will measure the situation over time, and will be bound to assume that improvements (if there were any) are attributable to the global impact of total monies spent, and not attributable to specific interventions. In this case, however, no practice (neither the 'best' nor the 'worst') can be considered responsible for the impact. This is, by the way, the alternative that fits best with the institutional arrangements that we have described above.

If, on the contrary, the evaluator considers that what is important is to know how impact has been attained and why, s/he is bound to consider that means (what Vedung [1998] calls the policy instruments) are relevant. Evaluation is then concerned with different ways of reaching objectives, and tries to judge which policy instruments, in isolation or in combination, and in what sequence, are better suited to the actors situated in given contexts.[20] It will not be possible to identify one single linear chain, or establish best practices. Rather, global policy impacts will be the result of the mix of policy instruments, to be accounted for in their combination and sequence.

Take the example of services for employment. The hierarchy of objectives could be seen to work as follows. The European objective is to raise the employment rate; the state passes a law for the creation of new Employment Centers;

local agencies deliver 'orientation' services to the unemployed based on knowledge of the labour market. The logical diagram of expected results, in its turn, works as follows. These services are supposed to improve the chances of being employed (employability), in order to create a better matching between demand and supply, which in its turn should favour employment. But will it operate the same way everywhere? Since each location has a different labour market (tight or loose; a balance of manual/clerical jobs offered or demanded; the presence/absence of an informal or irregular economy), the impact of raising the employment rate can be brought about in different ways. There will be a combination of activities and implementation pathways, and different 'partnerships' will be created in each case, with more or less tight links. What is at stake at the global level, is the appropriateness of the mechanisms utilized in the different contexts to reach consistent effects that can fulfill the global objective of raising the employment rate. So, unless the evaluation works out a theory of why something works better or worse in any given context, and the opposite in other contexts, it is meaningless to aggregate the global data of employment and declare the programme a success or failure depending on whether impacts were close to expectations. Nor is any best practice identifiable in terms of how a center is organized, or of whether it fulfils legal requirements, as if all the situations were alike, only some better and some worse along the same dimension.[21]

This leads us to models and best practices that are proposed as ways of generalizing local results across the board: in other words, what is proposed is to do the same things everywhere. There is a misplaced belief in 'best' as an absolute concept, one propose for all, while what we would need is something 'better', i.e. relative to particular situations (See Patton, 2001; Perrin, 2003). Furthermore, what is proposed is some 'thing' (an activity or a product), that can be compared, and not the actions and reactions that any situated actor may deem appropriate to a given mechanism offered by a programme. Instead, we need to compare different situations in order to understand why what fits in one case does not fit elsewhere, and vice versa. Such comparison, of course, would allow for better understanding of existing alternatives and the replication of examples that might appeal to the situation and favour flexible adaptation.

Take, for instance, micro-credit as a way to alleviate poverty. Many are now familiar with the Grameen Bank of Bangladesh (Yunus, 1999), and micro-credit has become a recipe for poverty alleviation everywhere. But it is rarely acknowledged that what made for the success of that experience was that poor Muslim women, working at home, were more responsible than their husbands, and that their morality pushed them to repay the debt, thus putting in place a real credit system, based on trust and co-operation. Where these conditions do not hold, micro-credit will be considered yet another subsidy, that people react to in a different way according to their own contexts. In those circumstances, other kinds of devices should be envisaged to foster trust and co-operation, before micro-credit is introduced.

In summary then, logical diagrams and best practices are attempts to utilize a theory of action that falls short of coping with the complexity of multi-level governance, because they do not address the crucial element of multiple means

to reach given goals. For this purpose, we propose instead to turn to theory-oriented evaluation. Weiss's theory-based evaluation offers to test the many theories of different implementers and stakeholders in similar situations. This could help investigate the way means are chosen and implemented in order to attain the desired goals. Pawson and Tilley's matrix of middle-range theory suggests finding a theory that can cope with the different outcomes of similar mechanisms. We refer in particular to the authors' concept of 'realistic cumulation' (Pawson and Tilley, 1997: 117), now expanded in Pawson's 'realist synthesis' (Pawson, 2002). According to the latter, evaluation should not set to itself 'the task of discovering whether a set of programmes works and 'aggregating' the results', rather its task is to 'test, refine and adjudicate the middle-range theories'. It is a process of abstraction, by which 'we move from one specific empirical case to a general theory and back to another case, and so on. What are transferable between cases are not lumps of data, but set of ideas' (Pawson and Tilley, 1997: 120). Sticking to our examples, evaluation should give to itself the task of demonstrating what employability mechanisms work in different ways for different labour market situations, how credit mechanisms work for different constituencies, and how different policy instruments could be combined for such situations.

It is clear that all this requires that the various institutional evaluation levels communicate among each other. If the institutional levels of evaluation do not communicate, it is impossible to relate outcome to impacts, or lower level programme results to higher level goals. If, however, they do communicate, it will be clear that many more alternatives exist than linear, virtuous, best practice, and evaluation has to account for them all.

Evaluating the 'Urban' programme

In this section, I will present a personal experience of the evaluation of a programme implemented inside a system of multi-level governance. It is a story of a voyage toward theory-oriented approaches, and a realization that for each step taken at the lower level, a counterpart at the higher level was needed and even now is missing.

The 'Urban' programme is a European Community Initiative, which promotes complex programmes similar to what are called 'Comprehensive Community Initiatives'. It is aimed at reducing social exclusion in urban areas by integrated projects of social inclusion in such different areas as urban infrastructure, health, education, employment and citizenship. It is run by the state, and money is allocated to local government which chooses an area among those eligible by reason of high scores on indicators of exclusion and which then contracts-out projects which may be urban, physical, economic or social (including education and health).

Having worked as the co-ordinator of a project called 'observatory of social initiatives'[22] in the Urban programme implemented in Rome (in the Tor Bella Monaca neighborhood), I and my colleagues started asking how the various projects fitted into the whole programme, i.e. how specific project objectives related to global goals. As we used to ask 'what has the toy-library got to do with

social inclusion?'.[23] Indeed, we spontaneously started with the 'hierarchy of objectives' framework in mind.

To our dismay, we soon realized that nothing of the kind was expected in the evaluation architecture on which the programme insisted.[24] The risk of vertical non-communication was serious. Being located at the project level, we were expected to deal only with outputs (activities, courses, meetings, etc.). There was surprise when we started asking the implementers questions that referred to the hierarchy of objectives ('why are you doing what you do?' 'What happens when you do it?' 'What is it that you consider a good result?') or to the integration between same-level projects ('what do you do best in co-ordination with other actors, that you could not do alone?'). At the beginning most implementers thought they were right in just being interested in running their toy-library, their laboratory on school attendance, their day-centre for disadvantaged people. They did not consider they were contributing to social inclusion. After all, according to the principal/agent theory they work in their own interests, why should they bother with what the programme wants to attain?[25] They expected to be evaluated for their own work with their own beneficiaries, not for the co-operative effort nor for contributing to the area's social inclusion. If they were carrying out their (local) 'practice', why should they bother with any (higher-level) 'theory'?

We decided to exploit as much of the knowledge that we were extracting from the local situation as we could, following theory-oriented approaches. We followed Weiss in distinguishing between implementation theory and programmatic theory. We realized that the operators might be familiar with implementation theory (which speaks to their activity) but not with programmatic theory (which speaks of the results of that activity), although they had many ideas about the latter. So, we started questioning the programme theory as it was written in documents, and we compared it with the theories held by implementers and stakeholders, and with what we ourselves were beginning to think.

The 'official' Urban programme theory offers two examples of logical diagrams of expected impacts. The first is a typical black box programme, indeed a 'complex black box'. Social exclusion is defined as a multi-dimensional problem because poverty overlaps with educational deficits, with personal disadvantage, drug addiction, etc. Hence, integrated programmes are called for to attack simultaneously the same area from multiple perspectives. The second is a mono-causal explanation: in poor neighborhoods, the cost of services is so high that local municipalities have withdrawn from them so that firms went away and jobs disappeared. If services were reintroduced by the programme, jobs will again be available thus recreating conditions for a healthy life.[26]

These pieces of theory, however, did not seem to properly account for what was going on. First of all, to describe the situation. It was clear that social indicators did not offer a fair picture of the area, which was very diverse. It was true that in some high-rise public housing, where it had been easier to get in if a handicapped person was a family member, there were concentrations of disadvantage. But other areas were better off, and the real challenge was how to bring the different sections of the neighborhood closer. Some projects were aimed at one

or other part of the population, and this might have aroused the jealousy of the part not involved.

Then, it was not clear how converging attacks on social exclusion could be successful, nor what they could produce. Every implementer was thinking of his/her own small target. However, each of them had ideas about what they were doing, and many other insights came out during the interviews and the focus groups we held with them. There was a clear understanding that collaboration between different groups was indeed possible, even if it was not taking place. Collaboration could work well under certain conditions: among third sector implementers only in very specific and serious cases; between local public servants (in the schools or in the local health agency) and third-sector implementers in cases in which a specific competence was needed by local public servants (e.g. in deciding whom to admit to day-centres). During the evaluation, people felt confident to develop their ideas about networking and collaboration, starting from what they thought were the positive points of their own actions.

As for the 'monocausal theory', it was simply ignored. There was no speculation about jobs returning in the future, as a path to social integration. Indeed, this piece of theory did not fit the local situation. There had never been jobs to loose, or – more likely – there were many irregular jobs, not accounted for by the official theory.

The prospect of a socially integrated future (the expected impact), however, is a more serious question, because it has to do with the long term, but everybody saw that there would not be any long term: the programme lasted three years, and it was clear that after this period most of the services would disappear. So we started inquiring about sustainability of the programme which led to a further surprise. When something was seen as a good outcome, people did not want to loose it. They organized in order to keep it going, sometimes in a co-operative way. We felt this very important in developing theories of empowerment.

We have come to fill the black box of the programme in a participatory way, by identifying ways of obtaining synergies, and mechanisms that had worked at least in a single project.[27] Among these mechanisms were:

- doing instead of getting: if a climate of trust is created between people, civil servants or NGO operators, then people do not blame absent services, but work together with the public sector to make them work (e.g. the employment centre);
- learning basics and trying it alone: people are taught basic technicalities, and then apply it to whatever situation they see the need (e.g. the rights centre)
- favouring mediation: working in group helps develop mediation capacities in places where individualism and conflict are endemic (e.g. the centre for disabled people and the multi-cultural educational laboratory)
- being pulled along: when something is felt to be good, then other people not directly involved want to take part in it.

Now, turning to the evaluation architecture, what will happen to these

findings? Will they be utilized in evaluating the whole programme? The terms of reference required that they be diffused (which has been done, in a proper publication: Urban Sottoprogramma Roma, 2001), but this does not mean that they will be assimilated in any way, nor that there is any guarantee that they will climb up in the programme's evaluation ladder.

At the level of national government, evaluation is managed by the Ministry of Public Works which has contracted it out to an independent evaluator.[28] Regardless of the so-called 'integrated policy', this is because public works of renovation and infrastructure building take the lion's share of these programmes. At the EU level, where the goals are stated and resources are allocated to states, there will be an ex post evaluation, undertaken by independent evaluators. What will it do? Assess the state of indicators of exclusion/inclusion? Count new jobs and consider them as an indicator of a good level of living, as the official programme theory would suggest?

Data on course attendance, number of actions taken, or square meters of public space improved, taken from various sites in Italy may be added up; some 'best practices' may be described (indeed even from Tor Bella Monaca there will be some examples). However an evaluation deficit is the least that can be expected if the evaluation system does not incorporate the most valuable results that can be found on mechanisms of improvement in collective life, and in collaboration between the public sector and the third sector.

Conclusion: How Can Theory-based Approaches Help Remove the Evaluation Deficit?

While theory-based approaches have helped open the black box of 'complex' programmes at a community level, programmes enacted through multi-level governance seem to suffer from a double kind of evaluation deficit: on the one side, under-utilization of lower level evaluations and, on the other, impact assessments unable to offer an explanation of why and how outcomes occurred.

This situation stems from the difficulty of reconciling a system of planning according to which the higher level seeks 'goals not means' with the search for linear causality chains and 'best practices'. Indeed, 'stating goals not means' is a way of coping with the complexity of reality, and of allowing each context to fully exploit its own abilities to move toward the accomplishment of global goals. Evaluation should match this by comparing the use of various means and assessing the way they worked in different contexts.

Theory-based evaluation approaches offer some clues to address this set of problems. Considering the beneficiaries of programmes as actors situated in a stratified reality in which they will act and react, theory-based approaches allow for different ways of conceiving how particular means and policy instruments will work to produce good outcomes. The latter should not therefore be seen as the result of the homogeneous implementation of some practice, nor should they be added up, to establish who fared well, who fared average and who fared the worst. Instead, the higher policy level should have the ability to recognize the potentialities of the lower levels, and of admitting to a multiplicity of means to be adapted to any given situation.

But the adoption of approaches like these would be meaningless, if institutional evaluation hierarchies like the one existing for evaluating Structural Funds programmes were not rethought. Given the separation between evaluation levels, theories emerged in the delivery of a programme do not bear on the understanding of effects and impacts up the ladder. It is more than a simple lack of communication. It is as if what actors have decided to do were not relevant to understanding how global impacts are achieved.

Notes

An earlier draft of this article was presented at the Fifth Conference of the European Evaluation Society at Seville. I have benefited from the comments that were offered there, especially by Elliot Stern, Petri Virtanen, Mauro Palumbo and Tassos Bougas. Thanks are also due to two anonymous referees who generously commented.

1. I take this definition from the programme of a conference on evaluating local development organized by OECD/LEED, Vienna, November 2002.
2. Scriven originally raised this point. His pragmatist approach proposed goal-free evaluation in which a judgement was reached on criteria decided upon by stakeholders with the evaluator's contribution (see Scriven, 1993). On the fact–value dichotomy, see House (2001: 313).
3. Of course, not all evaluations of the method persuasion were of an experimental type; more often they were simply descriptive, and limited to some technical device. But although rarely achieved, the experimental model remained the standard.
4. However, one could contrast Chen and Rossi's criticism with Campbell's idea of discarding 'plausible rival hypothesis' (Campbell, 1988).
5. There are mixed feelings about evaluation among policy analysts. Some policy analysts (Regonini, 2002) feel that evaluation can only account for the 'top down' approach and would reserve the 'bottom up' approach for themselves. Other policy analysts, among them the spokesmen for the implementation approach, include evaluation in the process of implementation (see Wildavsky, 1993: 213; Mazmanian and Sabatier, 1983: 9).
6. For an interesting debate about this, see Snider (2000). I discussed this in Stame (2001a).
7. I use the term 'theory-oriented evaluation' as a conventional way to group the three approaches I am considering, and I leave other expressions to those who invented them: 'theory-driven evaluation' to Chen and Rossi, 'theory-based evaluation' to Weiss. For an account of different evaluation approaches see Stame (1998) and Stame (2001b).
8. However, Scriven would not like to be considered a precursor to theory-oriented evaluation (see his ideas on 'minimalist theory' [Scriven, 1998]).
9. See how Chen and Rossi (1989) refer to this lack in Guba and Lincoln.
10. Note that Weiss introduced the utilization issue by maintaining that the 'utility test' should be as important as the 'truth test'; and once the centrality of utility had been grasped, she has indicated the importance not only of instrumental use, but also cognitive use.
11. See how Pawson (2000) evaluates a programme of training for inmates, utilizing Merton's middle-range theories of reference groups.
12. Weiss is a contributor to the Aspen Roundtable (Weiss, 1995).

13. Kubisch (1995: 4) calls this relationship a 'vertical complexity'.
14. In the Urban programme, we found this aspect notably absent, and we suggested greater attention to it in the future.
15. See O'Connor (1995) for an excellent review of the Comprehensive Community Initiatives programmes.
16. This problem refers also to other arenas of multi-level governance, like the development strategy of the Comprehensive Development Framework, that is the master plan of the World Bank (see Hanna and Picciotto, 2002). It states that in a given country the co-ordination is sought of various agents (international, national, local) acting in different sectors: it is not a division along the lines of level/competence, but a special mix that is created in any situation, where responsibilities are shared.
17. The concept of 'active subsidiarity' has been utilized in the experience of local development policies in Italy (Meldolesi, 2003). An analogous concept is that of 'mutual responsibility' worked out by Behn (2001).
18. See Coleman (1990), and how it has been elaborated for evaluation purposes by Virtanen and Uusikyla (2002) in a paper presented at the EES conference in Seville.
19. The MEANS guides (MEANS, 1999) are confusing on the relationship between the two logical schemes; on p. 94 it states that the objective tree cannot be used to identify causal links, whereas on p. 148 it states that the logical framework describes the sequence of effects, and that the column of statements of expectations is a theory of action. On this discrepancy, Palumbo (2001: 219–23) maintains that the objective tree and the causal tree do not live in the same forest.
20. 'How one thing leads to another', to follow Hirschman's strategy of economic development (1958).
21. This is, however, what has been mainly researched in the monitoring of Employment centres in Italy, see ISFOL (2002).
22. The Observatory was run by a group of researchers of the Department of Social Research of the University of Rome-La Sapienza. We intended to evaluate, and not simply to monitor, as the name would suggest. We were lucky to find agreement on this from the local office of the municipality who had commissioned our work.
23. The toy-library is a place where children spend their leisure time playing, under the supervision of a team of co-ordinators. It is considered a good initiative for how it is run, and for the novelty of such a thing in a place where children are usually unsupervised, either staying at home watching television or playing in the streets.
24. I am not raising here the problem of how projects were selected, and I am not assuming that some projects should not have been selected, although this was a hot political issue, especially among those who had been kept out. I want to make it clear that whatever project is selected it has to show its contribution to multi-level governance; and there must be a way of theorizing what this contribution is like.
25. This is why implementers fear the evaluation and accept control. They deal with the known, i.e. what they agreed to do, while evaluation may ask for the unknown, i.e. what might be expected. However, those who understand that evaluation can help them develop under utilized resources do not ignore this opportunity.
26. See the Terms of Reference for the programme Urban (94/C 180/02), especially section 1: 'Scope and Objectives'.
27. In this evaluation, we have thought of mechanism in the way Carol Weiss uses it (see above). At the time, we were less aware of the link between mechanism and context, as in realist evaluation. This could have helped better accounting for outcomes of projects implemented in the different sections of the neighbourhood. For a fuller discussion on mechanisms, see Hedstrom and Swedberg (1998).

28. See Università di Roma 3 (2002); this is mainly devoted to the physical side of urban regeneration and very little to integration among projects. The national evaluators have followed a top-down, hierarchy-of-objectives approach; they have commented on our activity, but have not taken into account our evaluation results.

References

Behn, R. (2001) *Rethinking Democratic Accountability*. Washington, DC: Brookings Institution Press.

Campbell, D. T. (1988) '"Degrees of Freedom" and the Case Study', in D. T. Campbell (1988) *Methodology and Epistemology for the Social Sciences: Selected Papers*. Thousand Oaks, CA: Sage.

Chen, H. and P. Rossi (1989) 'Issues in the Theory-driven Perspective', *Evaluation and Program Planning* (12)4: 299–306.

Coleman, J. (1990) *Foundations of Social Theory*. Cambridge, MA: Belknap Harvard.

Connell, J. P. and A. Kubisch (1995) 'Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiative', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss (eds) *New Approaches to Evaluating Community Initiatives*, vol. 1. Washington, DC: The Aspen Institute.

Cronbach, L., S. R. Ambron, , S. M. Dornbusch, R. D. Hess, R. C. Hornik, D. C. Phillips, D. F. Walker and S. S. Weiner (1980) *Toward Reform of Program Evaluation*. San Francisco, CA: Jossey-Bass.

Etzioni, A. (2001) 'Humble Decision Making', in P. F. Drucker, R. L. Keeney and J. S. Hammond (eds) *Harvard Business Review On Decision Making*. Boston, MA: Harvard Business School Press.

EU (2000) *Structural Funds 2000–2006*. Luxembourg: Office for Official Publications of the European Community.

Guba, E. and Y. Lincoln (1989) *Fourth Generation Evaluation*. Newbury Park, CA: Sage.

Hanna, N. and R. Picciotto (2002) *Making Development Work*. New Brunswick, NJ: Transaction Publishers.

Hedstrom, P. and R. Swedberg (1998) 'Social Mechanism, an Introductory Essay', in P. Hedstrom and R. Swedberg (eds) *Social Mechanisms*. Cambridge: Cambridge University Press.

Hirschman, A. O. (1958) *The Strategy of Economic Development*. New Haven, CT: Yale University Press.

House, E. (2001) 'Unfinished Business: Causes and Values', *American Journal of Evaluation* 22(3): 309–15.

ISFOL (2002) *Servizi all'impiego e decentramento*. Milano: Angeli.

Kubisch, A. C., C. Weiss, L. B. Schorr and J. P. Connell (1995) 'Introduction', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss (eds) *New Approaches to Evaluating Community Initiatives*, vol. 1. Washington, DC: The Aspen Institute.

Mazmanian, D. A. and P. A. Sabatier (1983) *Implementation and Public Policy*. Glenview, IL: Scott, Foresman & Co.

MEANS (1999) *MEANS Collection – Evaluating Socio-economic Programmes*. Luxembourg: Office for Official Publications of the European Communities.

Meldolesi, L. (2003) 'Nuovi lavori, vecchi lavori: una nota duale', in *Atti del Convegno in memoria di Marco Biagi 'Verso uno statuto dei lavori?'*, Perugia, 8 Novembre 2002. Napoli: Edizioni Scientifiche Italiane.

O'Connor, A. (1995) 'Evaluating Comprehensive Community Initiatives: A View from History', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss (eds) *New*

Approaches to Evaluating Community Initiatives: Volume 1, Concepts, Methods, and Contexts. Washington, DC: The Aspen Institute.

Palumbo, M. (2001) *Il processo della valutazione*. Milano: Angeli.

Patton, M. Q. (2001) 'Evaluation, Knowledge Management, Best Practices, and High Quality Lessons Learned', *American Journal of Evaluation* 22(3): 329–36.

Pawson, R. (2000) 'Middle-range Realism', in *Archives Européennes de Sociologie* 41(2): 283–324.

Pawson, R. (2002) 'Evidence-based Policy: the Promise of "Realist Synthesis"' *Evaluation* 8(3): 340–58.

Pawson, R. and N. Tilley (1997) *Realistic Evaluation*. London: Sage.

Perrin, B. (2003) 'How Evaluation Can Help Make Knowledge Management Real', in R. Rist and N. Stame (eds) *From Studies to Streams*. New Brunswick, NJ: Transactions Publishers.

Regonini, G. (2002) *Le politiche pubbliche*. Bologna: Il Mulino.

Scriven, M. (1993) *Hard Won Lessons in Program Evaluation*, *New Directions in Program Evaluation*, 58. San Francisco, CA: Jossey-Bass.

Scriven, M. (1998) 'The Least Theory that Practice Requires', *American Journal of Evaluation* 19(1): 57–72.

Snider, K. F. (2000) 'Rethinking Public Administration's Roots in Pragmatism', *American Review of Public Administration* 30(2): 123–145.

Stake, R. (1980) 'Program Evaluation, Particularly Responsive Evaluation', in W. B. Dockrell and D. Hamilton (eds) *Rethinking Educational Research*. London: Hodder and Stoughton.

Stame, N. (1998) *L'esperienza della valutazione*. Rome: SEAM.

Stame, N. (2001a) 'La cultura della valutazione, tra pragmatismo e istituzionalizzazione', in F. Battistelli (ed.) *La cultura delle amministrazioni pubbliche tra retorica e innovazione*. Milano: Angeli.

Stame, N. (2001b) 'Tre approcci principali alla valutazione: distinguere e combinare', in M. Palumbo (ed.) *Il processo della valutazione*. Milano: Angeli.

Sullivan, H., M. Barnes and E. Matka (2002) 'Building Collaborative Capacity through "Theories of Change". Early Lessons from the Evaluation of Health Action Zones in England', *Evaluation* 8(2): 205–26.

Urban Sottoprogramma Roma (2001) *Valutazione degli interventi a carattere sociale*. Roma.

Università di Roma 3 (2002) *Valutazione del programma URBAN, Roma*. Roma: Università di Roma 3, Facoltà di Architettura.

Vedung, E. (1998) 'Policy Instruments: Typologies and Theories', in M. L. Bemelmans-Videc, R. C. Rist and E. Vedung (eds) *Carrots, Sticks and Sermons: Policy Instruments and their Evaluation*. New Brunswick, NJ: Transaction Publishers.

Virtanen, P. and P. Uusikyla (2002) 'Exploring the Missing Link between Causes and Effects'. Paper presented at the European Evaluation Society conference, Seville, Spain, 12 October.

Weiss, C. (1987) 'Where Politics and Evaluation Research Meet', in D. Palumbo (ed.) *The Politics of Program Evaluation*. Newbury Park, CA: Sage.

Weiss, C. (1995) 'Nothing as Practical as Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives for Children and Families', in J. P. Connell, A. C. Kubisch, L. B. Schorr and C. H. Weiss (eds) *New Approaches to Evaluating Community Initiatives: Volume 1, Concepts, Methods, and Contexts.* Washington, DC: The Aspen Institute.

Weiss, C. (1997) 'Theory-based Evaluation: Past, Present and Future', in D. J. Rog (ed.)

Progress and Future Directions in Evaluation, *New Directions for Evaluation*, 76. San Francisco, CA: Jossey-Bass.

Weiss, C. (2000a) 'The Experimenting Society in a Political World', in L. Bickman (ed.) *Validity and Social Experimentation: Donald Campbell's Legacy*. Thousand Oaks, CA: Sage.

Weiss, C. (2000b) 'Which Links in Which Theories Shall We Evaluate?', in P. J. Rogers, T. Hacsi, A. Petrosino and T. A. Huebner (eds) *Program Theory in Evaluation: Challenges and Opportunities*, *New Directions for Evaluation*, 87. San Francisco, CA: Jossey-Bass.

Wildavsky, A. (1993) *Speaking Truth to Power*. New Brunswick, NJ: Transaction Publishers.

Yunus, M. (1999) *Il banchiere dei poveri*. Milano: Feltrinelli. (*Vers un monde sans pauvreté* translated by Jean-Claude Lattès.)

NICOLETTA STAME is Professor of Social Policy at the University of Rome 'La Sapienza'. She is a co-founder and former president of the Italian Evaluation Association (AIV) and is president of the European Evaluation Society (2004–05). Please address correspondence to: Dipartimento di Ricerca Sociale, Corso Italia 38a, 00198 Rome, Italy. [email: nstame@aconet.it]