WE1 Nicoletta Stame [*]

Evaluation and the policy context: the European experience[*]

My talk today deals with the European experience with evaluation. A few points are necessary to frame what I am going to talk about

First, I will be concerned with the development of a European tradition in evaluation, and I will consider two different instances of it: one, European evaluation as it was developed by various European countries; two, evaluation as it became understood inside the European Union context. In both cases it is a matter of a diverse and multicultural political context.

Second, I will explore my topic against the international experience of evaluation, of which the Australasian one is an important part.

Third, I will try to see how the European experience can contribute to an understanding of your theme, namely the politics of evaluation.

Fourth, I will help myself with the periodization of evaluation dissemination that was worked out in the *International Atlas of Evaluation* (ed. by Ray Rist, J.E. Furubo and R. Sandhal); the latter considers 25 countries, including Australia, New Zealand, and Italy. That periodization identifies three phases and two great causes for the bringing about of evaluation: social programs and the public sector reform known as New Public Management (NPM).

The three main phases are the following:

1. the '60s – '70s: when in the USA the first programs of the Great Society were launched, that had evaluation mandated, according to the sunset legislation. In this period only Canada joined the US[1]
2. the '80s: when some countries of the Anglo-saxon tradition started introducing NPM reforms of the public sector. Here the UK, Australia and New Zealand are foremost, and some Northern European continental countries follow suit.
3. the '90s: when evaluation was extended to many more countries, thanks to an external push, coming from larger agencies, like the EU for many European countries , or the World Bank or other international agencies for such Third World countries as China, Korea, Zimbabwe. In this period the evaluation community has grown internationally, creating networks and links that make for a much greater communication than in the past..

Old Europe and evaluation

Speaking in New Zealand, a "new" country where you are celebrating the 21st anniversary of the AES, it is easy to think of ourselves through the stereotype of "old Europe". Indeed, from the

[*] University of Roma "La Sapienza", Italy; vice-president of the European Evaluation Society

[*] speech delivered at the Conference of the Australasian Evaluation Society, Auckland, 17 september 2003

[1] in fact, there are European situations that pretend to have established evaluation at this time: Germany, with its Lander system that imitates the American federalism, a strong facilitator of the first social experimentations; Sweden, that imitated the US with its program of "The strong society" (Vedung, 1997, p. 27).

point of view of evaluation Europe has not been an innovator: and our EES has been founded only in 1995. Why did evaluation come so late to our shores?
I can see three main reasons:

a. a strong ideological tradition in the state intervention in socio-economic affairs: the welfare state was born here. In Europe there could have been nothing like the debate on state intervention that accompanied the introduction of the War on Poverty, and that gave way to evaluation. In Europe nobody doubts (or at least did not doubt until recently) about the need for the public provision of social services and for the implementation of public policies: the latter are goods in themselves, that do not need being evaluated. The highly ideological debates refer to the amount of social spending (inputs) rather than to outputs or outcomes.
b. An administrative tradition, at least in continental countries that have an administrative law, according to which the public servant is judged for the legitimacy of his/her acts, not for their results. Procedure comes first, outcomes come if they may.
c. A strong presence of TUs of the public sector that are more interested in workers' rights (that they feel threatened by evaluation!) than in citizens' needs.

This mix has produced a malfunctioning of the state and at the same time an inability to reform.

This situation has evolved in the '90s. The UK, the home country of lord Beveridge and of the universalistic model of the welfare state, has moved to public sector reforms that have been adopted also in other continental countries, with various ideological blending (from the French left wing government of Rocard to the Italian center-left governments of the '90s, to other conservative governments). In the same period, the evaluation community of the other countries had come to maturity, and this had made room for dialogue and debate among those people in Europe who became concerned with evaluation. It was clear that we could learn from each other. The European Evaluation Society was founded in 1995.

In the following sections I will try and understand how has Europe participated in the development of evaluation worldwide.

The first wave: social programs

Program evaluation started in the US with the Great Society. What was created then were new political tools, new ways of acting for tackling social problems, and a new practice, that of evaluation.

Just to remind some main points briefly,

- programs can be understood as actions for the purpose of obtaining a change, they have to be implemented with given means in order to obtain intended results within a given deadline.
- they require to be evaluated in order to know whether they were effective. Program evaluation is a field to which both Michael Patton and Michael Scriven (keynote speakers in this Conference) have contributed in a paramount way.
- a method for evaluation had to be worked out. Some thought it was just a matter of methodology, but things were not as easy: positivists and constructivist paradigms opposed each other: the former proposing variable analysis (along the teaching of the Bureau of Applied Social Research) and experimentalism (Campbell), the latter proposing qualitative analysis and actor's involvement (Stake's responsive evaluation, Guba and

Lincoln's fourth generation evaluation). And Patton tried to tame the "methodological dragon". Later on (in the late '80s) this problem was addressed by way of multi-method approaches (Greene and Caracelli). Others spoke of a logic of evaluation (Scriven).

- Since the beginning, evaluation has been flawed by the "black box" problem: programs were conceived as a causal sequence, but often without explaining why a certain result should be the effect of a program seen as its cause. First attention to this problem was given by Chen and Rossi (1989) , and especially by Carol Weiss (1997).

The main legacy of this period is therefore twofold:

- the paradigm debate should be mastered by some form of pluralism of methods
- the black box problem needs an elaboration, that further on has taken the form of the theory based approaches.

Nothing of the kind was present in Europe at this time. The '70s has been a decennial of great "structural reforms" (in Italy a reform of the Health system, of social security): all the hopes rested on the political decision, and there was an assumption that implementation would follow. At the same time, programs would have been considered partial, fragmentary, not up to the situation.

Not having participated in the first wave, and in its debates, the European evaluation community is thus not aware of the sufferings of the beginning. So, when it later entered the field there was an expectation of finding ready made methods and techniques. Analogously, it was not aware of the need of avoiding certain errors, that were reiterated (one is not vaccinated against committing the same errors). For instance, there was scarce knowledge not only of the first paradigm wars between the positivist and constructivist approaches to evaluation, but also of their overcoming through Patton's sweeping critiques of the methodological dragon, though multimethod approaches and the "paradigm of choice" (Patton, 1986). This means that we are still waging a rearguard war between qualitative and quantitative methods. This has been reinforced by the EU, that initially imposed a strong quantitative imprint, while the qualitative methods, for all the discourse on methodological pluralism, still have to fight for being accepted.

The second wave: New Public Management

At the beginning of the '80s, with a growing fiscal crisis of welfare states, new problems emerged in industrialized states that had expanded social policies: that of reducing public spending.

A new thinking of the role of the state emerged, that put different questions to evaluation. It is the idea of the New Public Management, firstly developed in Australia and New Zealand, but also in some European countries (Pollitt and Bouckaert, 2000).

The characteristics of the NPM can be summarized as follows:

- there is a change in the role of the state, "steering, not rowing" (Osborne and Gaebler, 1992). Not all state functions are alike: the specific state function is to lead, to give orientation, not to execute
- therefore, many functions can be outsourced to private or third sector agencies.
- What regulates this relationship is the "principal/agent" concept, according to which there are distinct roles and responsibilities between the "principal", or the state agency, which is interested in the outcome of an intervention, and the "agent", or the contractor, who is responsible for the output of an action.

This system puts new problems for evaluation:

- new criteria of evaluation become prominent, beyond effectiveness: see in particular the UK elaboration of the concept of "value for money", meaning that there must be a direct return for any penny spent. In practice, this has meant a growing attention to efficiency than before.
- A new distinction becomes usual currency: that between accountability and learning. Both are new concepts. While one could say that a knowledge function was present since the beginning, owing to the social science background of the first evaluators, the learning function draws the attention toward the ability of actors and stakeholders to benefit from the evaluation. On the other hand, having identified the agent more clearly, s/he is held accountable to the principal for his/her deeds.

The main problems that have been raised vis-à.vis the new phase can be seen as the following:

First problem. Stemming from a corollary of the principal/agent concept is the separation between the leading role of the principal and the executive role of the agent. This separation can be interpreted under many different lights, that I would distinguish between its "adversarial" and "collaborative" aspects.

It unfolds its **adversarial** aspects when principal and agent behave as two separate agencies, having conflicting goals. On the one hand, the principal aims at getting a given result by the simple way of disbursing a sum, without having to engage in a defatigating administrative tasks. On the other hand, the agent aims at keeping the contract going, and the money coming in, often being more interested in the maintenance of his/her own resources than in the program goals[2].
As a consequence, the following negative repercussions on evaluation can be detected:

- departments or contractors (agents) feel responsible only for the output they are expected to deliver, not for how it leads to outcome. A consequence of this is that agents easily accept monitoring of their activity (which is what they have contracted), but not evaluation, that would call them to test for something they feel not being in their command. And they will develop strategies of resistance to evaluation.
- Ministries (principals) are mainly interested in the act of contracting out, since they leave the action to the agent, hoping that a good contract will release a good product. As for evaluation, they praise ex ante evaluation, and are not interested in ex post evaluation, often airing the false assumption that a good specification of requirements in contracts leads to good results, in the same way as the program designers' bias that a well designed program will bring good results.

The combined result of these two tendencies is an actual undermining of evaluation.

However, that same relationship could unfold **collaborative** aspects.
This situation could be studied with Robert Behn's idea of "democratic, mutual accountability" (Behn, 2001), according to which a sharp separation between principal and agent can be counterproductive: the principal should be responsible for putting the agent in the condition of acting, the agent should feel responsible for the outcome.
In order to get this, it would be necessary to get principal and agent nearer:

- they should share goals, not have their own complementary (and in fact conflicting) ones. Both (not only the principal) should contribute to program elaboration, goal definitions etc.,

- they should understand that they can both benefit from evaluation, the principal by following the various phases of the intervention, and not only the ex ante evaluation; the

[2] This can be seen when agents win contracts for something they are well versed in, irrespective of the policy goals.

agent by considering process evaluation as an instance of empowerment for his/her capacity of addressing the present, and other future, situations in which he/she is engaged. [3]

Second problem. The principal/agent concept can only deal with single activities, but it is unable to account for mixed interventions, which are, however, the great majority of programs (Turner and Washington, 2002, p. 367)

What happened in Europe during this second wave?

- Some European states introduced systems of NPM, and evaluation as a consequence of this. In some cases this led to an exaggeration: see the "audit explosion" in the UK, as exposed by Michael Powell (1997).
- The EU introduced programs for the social and territorial re-equilibrium. The first such programs were called "Poverty" (reminiscent of the US "war on poverty"). Other important programs were the "Integrated Mediterranean programs", that have been pivotal for mandating evaluation.

Third wave: evaluation diffusion

During the '90s evaluation has become widespread all over Europe, with a mix of the two previous trends.

In almost every national state there were reforms of the public sector introducing some aspect of NPM (in France with the left-wing Rocard government, in Italy with the center-left government), meeting greater or smaller resistance. At the same time, the political systems became more open to working with programs, in the social, public health, employment, environment, education etc. sectors, that in due time became more and more complex, integrated, multidimensional. In these instances, you have a growing interest in getting methods and techniques from the outside, but also the development of original approaches. National characteristics are therefore worked out, with UK and Northern countries more linked to the anglo-saxon debate, and other countries building on their oen cultural traditions.

But what has been a crucial spurt to evaluation has been the external push coming from the EU Structural Funds (social funds for human resources and employment, for territorial reequilibrium and social cohesion, for rural development) that have represented a great mobilization of financial and human resources, and have required monitoring and evaluation of their results. This push has brought with itself a special evaluation style, that initially has had a greater impact on countries of the third wave of evaluation institutionalization, but that is at work also in countries of the second wave.

In fact, the EU has developed a complex system of multi-level governance that is a strange mix of social programs and principal/agent principles, of which a specific architecture of evaluation is a crucial element.

In order to understand how all this works, it is necessary to remember that the EU is a federal system, that has similarities and differences with other federal systems like the USA, Canada or Australia. As in the latter cases, there is a division of competencies among levels (in the EU we have: EU, state, region, municipality) , and a devolution to the lower levels of many tasks. What

[3] Talking about international development evaluation, Picciotto said that the partners (which is already a collaborative concept), i.e. donor agencies and beneficiaries, should "share objectives, have distinct accountabilities, have reciprocal obligations" (AJE, 2003, 232)

distinguishes the European experience, however, is the legacy of her history. The EU was born after centuries of wars between the states that now compose it and have decided to live peacefully: the European unity is a gradual process of unification, by which the single states give up pieces of sovereignty in order to build a new entity. Her diversity and multiculturalism [4] are constituting elements, assets to be maintained through an original model, that I would describe as follows.

The *policies* where it applies [5] establish great goals: territorial re-equilibrium and social cohesion (ob. 1 of FESR), an integrated rural development (ob. 5 of FEOGA), human resources development (along the axes of employability, entrepreneurship, flexibility and equal opportunities). All these big goals cover the multiple dimensions of reality.

The *decision-making process*.

- It is characterized by incrementalism: negotiations about conflicting interests among states, or states and the EU Commission (especially when local governments are of a different political side from EU commissioners)
- it is limited to financial contribution: the main conflict of interests hinges on allocations to the states
- it tends to keep in the many vested interests of beneficiaries, implementers, etc.[6]

This system of *multi-level governance* is characterized by the motto "the EU states goals, not means"; the latter are established at the lower levels, where the money is spent:

- the EU establishes general goals and allocates money to the states
- the states establish specific/intermediate goals and allocate money to the regions (or provinces, or municipalities, or to specific sectors)
- the lower levels decide about programs and intervention and here it is where the money is spent.

The rule that regulates the relationships among these levels is the principle of subsidiarity: the higher level does not do what the lower level can do.

An *evaluation logic* follows from all this:

- the EU is more interested in financial evaluation (how the money is spent) than in effectiveness of the interventions – at least this is what happened at the beginning
- an evaluation hierarchy has been institutionalized along the multi-level governance: evaluations have been mandated at the EU, state and local level in order to assess the correspondent level of spending. At the beginning it was mainly a matter of commissioning evaluations, now it is a matter of creating evaluation units.

Analogously with what happened with the principal/agent principle of the NPM, the European subsidiarity can develop adversarial aspects as well as collaborative ones.

There have been **adversarial** aspects when the lower level did not want the EU or the state to bother into their own affairs: a new regionalism and localism that fought for extreme devolution

But it is possible to develop **collaborative** aspects, when the top is concerned with helping the lower levels doing what they can do best, with latent resources, etc.: what we call "active subsidiarity".

---

[4] In europe, multiculturalism does not refer to a problem of integrating minorities: it is the very texture of the European society that is multicultural.

[5] not all policy domains are decided at the European level: how to enlarge the scope of the latter is the main issue.

[6] These groups, by the way, do not want to be checked by a non-political activity like evaluation.

The following table can illustrate the span of options open to the actors of this system of governance

	PRINCIPAL /AGENT	**SUBSIDIARITY**
adversarial	- principal responsible for outome, agent responsible for output - do not share goals: conflict of interests - agent accepts control, but not evaluation - principal is only interested in ex ante evaluation	- strict vertical division of competencies - lower level does not want higher to meddle with it: extreme devolution - top level only interested in financial control, not in effectiveness
collaborative	- principal and agent share goals - principal and agents are both interested in success - agent accepts evaluation of how output leads to outcome MUTUAL ACCOUNTABILITY	- the top is concerned with what the bottom can do (helps exploit latent resources) ACTIVE SUBSIDIARITY

Actual European predicaments for evaluators: remedies, alternatives

So far we have seen the premises (complexity of dimensions, plurality of levels), now let's look at the problems. What do evaluators do when the top level states goals not means?

According to the table above, one could see two alternatives.

One, in line with the adversarial mode. The top is only interested in assessing whether goals were achieved, not in how they were achieved (variety is admitted, but is not relevant):

- rough indicators of goals is what matters
- they are used in a pre-post verification logic, not even in an experimental one, because there is no identification of a program (experimental) situation vs. a non-program (control) situation.

The limit of this approach is that there is neither learning (little understanding of the process) nor accountability, because the link between the agent/implementer and the principal/EU are too loose.[7] Consequently, the higher level is a passive receiver of information.

Two, in line with a collaborative mode. The top is interested in understanding what works better, where and why. Therefore, it has to develop ways of understanding differences. In this case, it is constantly concerned with the lower levels, in line with the options of mutual accountability and

[7] A similar point is raised by B. Ryan (2003, p. 13) in an article on monitoring and evaluation in Australia and New Zealand

active subsidiarity. Here we would find a learning organization approach, in which all parties were involved in actively producing information for evaluation.

However, this is not the framework with which the EU utilizers of evaluation judge the evaluations that they receive up the ladder. They usually express a dissatisfaction with the quality of evaluations on the following accounts: data are of a low quality; evaluators are too politicized and not independent; evaluations are not clear about the program logic: ex ante evaluations are not good, the logic of action is missing, there is no evaluability assessment etc. . In other words, evaluations are considered inadequate because they neither perform the task of generalizing results (summative evaluations) nor offer suggestions for improvement (formative evaluations). All this is usually attributed to the complexity and ambiguity of programs that "seldom have well-specified or quantifiable objectives" and to the "poorly-developed monitoring systems" (Summa and Toulemonde, 2002, p. 417).

Two main remedies are proposed:

a) work with the *program logic*. Various models are proposed of the vertical links between general goals (at the EU level), intermediate goals (state levels) and operational goals (local, intervention level). This would allow to understand causal links between what happens at the various levels.
b) establish *best practices*. Identify the best application of planned interventions (actions, services, etc.).

These remedies to bad quality evaluations are at odds with the policy of "stating goals not means". Instead of proposing how to account for the complexity and variation that is implied in the latter, they consider them as an accident to be overcome through models and generalizations.[8] Taking up Patton's classification[9], we could say that what is proposed is a "linear" model of evaluation in a "systemic" situation.

My contention would not be that evaluations are not bad[10], but that if evaluations have to match the requirements of the multi-level governance of the EU, they have to take other paths to be relevant. These paths should fit the two main streams that have merged in the European evaluation tradition: social programs and NPM. And they would be an alternative to the proposed remedy.

Theory-based evaluation for complexity

The first alternative deals with the program logic: it refers to the social program tradition and to its developments. Take the MEANS guides, that distinguish between the "hierarchy of objectives" and the "logical diagram of expected impacts" (vol. 1, p.93 and 95).

- the former works top down, establishing a cascade of objectives: the result of the higher level are the goals of the lower one"
- the latter works bottom up. The assumption is: to get certain results you have to do certain things: if you do "a" then you will get result "b", that will have impact "c".

The logic behind these models is that there is only one theory of how things get done, the good theory; and that the program is articulated into a series of virtuous linear chains from the results of local intervention, to the effect on performance of national programs, up to impact of EU policies.

[8] The same happened with program evaluation in the US when Weiss, Cronbach, Patton, among others, put the policy context at the center of evaluation, against the positivist tradition that considered it as a "threat to validity".
[9] I refer to Patton's keynote speech delivered at the AES conference the day before this one.
[10] Perhaps one could see here a similarity with what B. Ryan says of the Australian situation: "Austrsalia wanted too much evaluation too quickly" (2003, p. 7).

However, to be in tune with the complexity problems of the EU multilevel governance, we could elaborate what I would call a "theory-based evaluation for complexity". I see two main instances of it:

- with Carol Weiss (1997)'s approach, we could say that among implementers of the great goals policies there are many theories: let's see what mechanism worked in a specific situation; let's ask it to stakeholders, implementers, etc.. This approach is likely to be more friendly toward the various stakeholders that implement, and benefit from, any European program.
- with Pawson and Tilley (1997)'s realistic evaluation we could say that the outcome depends on the combination between a given mechanism (incentives, regulations, persuasion, providing services or training) and the context[11]: how people embedded in different situations decide to use them[12]. The combination of mechanism-context –outcome will tell us "what worked better, where, in what circumstances and why".

While the hierarchy of objectives is separate from context, and assumes that a given tool always works in the same way, these theory-based approaches start from complexity of aspects and multiplicity of contexts, and assume that what will work is always a combination of tools[13].

I can take the example of the European policy of employment.
The EU goal is: raising the employment rate. The state goal is: improve labor supply/demand matching. The local intervention is: creating employment centers.

There are however various problems:

- context: each site has a different labor market (tight/loose; manual/clerical jobs are offered)
- tools: there are many tools for that goal; there can be a different combination of tools in each site.
- Theory: it should say why some tool works better in what context: e.g. where there are irregular jobs, services for surfacing should be necessary.

"Best" practices or "good" practices?

The logic of best practices is an attempt at establishing that all situations are alike (it is possible to generalize) , and that simply some actors are better than others. Invariably, some places are always better (for example: Emilia Romagna and some always worse (the South in general). If the latter are not up to the former it is their fault.

The contrary is true:

a) nothing can be considered best for all situations, hence generalizable
b) there are different situations, and something that has shown to be good somewhere perhaps could be studied and adapted/imitated somewhere else.

[11] Pone could guess that the great popularity of realistic evaluation owes much to its focussing on the difference of context, that is central to the European perception.
[12] Yesterday there was a comment against the idea of programs as "change agents" seen as something imposing change from above; in the realistic evaluation conception, on the contrary, programs are seen as opportunities that the beneficiaries may decide to take, hence as facilitators of a change that remains in the hands of actors.
[13] On this aspect, see Vedung and Salamon.

All this could be attained if there were a continuous interaction between lower and higher levels of the hierarchy, if good practices were reflected in good theories, and there were learning about it. But nothing of the kind is possible inside the existing institutional hierarchy of the EU system of evaluation, where

- lower levels are expected to do only monitoring, not real evaluation
- higher levels do not receive contributions from below.

c) such a link between theory and practice would allow for the enactment of a principle of learning organization, in which mutual accountability would prevail:
   - at the lower level, if people knew what they were doing, what to expect, how they could contribute to outcome, then they would be more favorable to evaluation, and understand that it is for their good
   - at the higher level, theories received from below would be more grounded, and better suited to understand a complex and diverse situation.

To conclude

Contrary to what happened up to ten years ago, the European scene is now an integral part of the international evaluation community. The single European countries may follow their particular paths, but at the same time they are influenced by a new governance model, the multi-level governance of the EU. And while all countries have known their particular ways of adapting program evaluation and NPM practices, the EU multi-level governance system, that is operating in all countries, is a particular mix of both traditions of social programs evaluation and NPM.

I have proposed that the European evaluation community tried to overcome its actual predicaments by:

- building on the evolution of program evaluation, through "theory based evaluation for complexity"
- building on the evolution of NPM concepts, especially its links with the learning organization and mutual accountability.

The whole international evaluation community faces problems linked to complex programs (that are an evolution of programs) and of multi-level governance (that is an evolution of NPM). I hope the European experience may be of some interest also to others, if it works as a benchmark, not as a model: we live in a global world, we share problems, we are different.

References

Behn, R., 2001, *Rethinking democratic accountability,* Washington, DC: the Brookings Institution

Chen, H. and Rossi, P., 1989, "Issues in the theory-driven perspective", in *Evaluation and Program Planning*, vol. 12, n. 4

Furubo, J.E., Rist, R. and Sandhal, R, eds., 2002, *International Atlas of Evaluation*, New Brunswick, NJ: Transactions Publishers

Furubo, J.E., and Sandahl, R., 2002, "A diffusion perspective on Global Developments in Evaluation", in Furubo, J.E., Rist, R. and Sandhal, R, eds.

MEANS collection, 1999, *Evaluating Socio-economic Programmes*, 6 voll., Luxemburg: Office for Official publications of the European Communities

Osborne, D. and Gaebler, T., 1992, *Reinventing Government*, Reading, Ma: Addison-Wesley

Patton, M., 1986, *Utilization-focused evaluation*, Newbury Park, Ca : Sage

Pollitt, C. and Bouckaert, G., 2000, *Public Sector Reform*, Oford: Oxford University Press

Powell, M., 1997, *The Audit Explosion*, Oxford: Basil Blackwell

Ryan, B., 2003, "Death by evaluation? Reflections on Monitoring and evaluation in Australia and New Zealand", in *Evaluation Journal of Australasia*, n. 1

Summa, H., Toulemonde J., 2002, "Evaluation in the European Union: addressing complexity and ambiguity", in Furubo, JE, Rist R. and Sandhal, R., eds.

Turner D. and Washington S., 2002, "Evaluation in the New Zealand Public Management System", in Furubo, JE, Rist R. and Sandhal, R., eds.

Weiss, C., 1997, "Theory-based evaluation: past, present and future", in D.J. Rog, *Progress and Future Directions in Evaluation*, New directions for Evaluation, n. 76, San Francisco, Jossey Bass