

Evaluation in public-sector reform: Toward a 'third wave' of evaluation?

Hellmut Wollmann

Humboldt-Universität zu Berlin

hellmut.wollmann@rz.hu-berlin.de

in.: Hellmut Wollmann (ed.) 2003, Evaluation in Public Sector Reform,

Cheltenham/Northampton: Edgar Elgar, pp- 1 - 11

I. Introduction: The ‘twinned’ development of public-sector reforms and evaluation as a ‘missing link’ in research and debate

The initial premise and thesis of this volume is this: Public-sector reform and evaluation have been closely interlinked like, as it were, Siamese twins throughout the past 30 years or so. Yet an inspection of the available literature on public-sector reforms and evaluation reveals a glaring discrepancy and gap: While the fields of public-sector reform and of evaluation have each brought forth a huge body of literature and research, these two realms have been largely treated as separate entities. Their ‘twin-like’ connection so far has received little attention.

This book aims to contribute to the bridging and filling of this gap. In addition, the volume covers more countries than most of the available publications on public-sector reforms¹. In addition to addressing the ‘usual suspects’ in the current international debate (that is, the Anglo-Saxon and Scandinavian countries), the volume contains articles on Continental European countries, Japan, and Latin America—16 countries in total.

II. The three phases of ‘twinning’ of public-sector reforms and evaluation

Roughly three phases in the development of public-sector reform and evaluation over the past 30 years can be distinguished: the first wave of evaluation during the 1960s and 1970s; the second wave beginning in the mid-1970s; and a third wave related to the New Public Management (NPM) movement.

During the 1960s and 1970s the advent of the advanced welfare state was accompanied by the concept of enhancing the state’s capacity for ‘proactive policy making’ through a profound modernisation of its political and administrative structures, for the pursuit of which the institutionalisation and employment of planning and evaluation capacities was seen as strategically important. Conceptually this was premised on a ‘policy cycle’ revolving around the triad of policy formation and planning, implementation and evaluation, whereby evaluation was deemed instrumental as a ‘cybernetic’ loop, gathering and feeding back information relevant to policy making. Policy evaluation, ideally conducted as full-fledged social science-based evaluation research, was primarily directed at the output and outcome of (substantive) policies. Embedded in the reformist mood (and optimism) of the (short-lived) ‘planning period’, policy evaluation was (in its normative sense) meant to improve policy results and to maximise output effectiveness. This early phase of policy evaluation has been called the ‘first wave’ of evaluation (for an early conceptualisation and interpretation, see Wagner and Wollmann 1986, Derlien 1990). While the US has been the global pacesetter of policy evaluation since the mid-1960s, in Europe, Sweden and Germany were the frontrunners in this ‘first wave’ of evaluation (for an early perceptive comparative assessment, see Levine 1981).

Since the mid-1970s, in the wake of a world-wide economic and budgetary crisis triggered by the first oil price shock of 1973, policy making has been dominated by the need for budgetary retrenchment and cost efficiency. Consequently, the mandate of policy evaluation has been redefined, in that the implicit task was to reduce policies and maximise input efficiency. From a developmental perspective, this phase was the ‘second wave’ of policy evaluation. Among the European countries, the Netherlands and Great Britain were exemplars of this wave (see Derlien 1990).

A third wave of evaluation came into being during the late 1980s and 1990s, with ever more pressing budgetary crises in many countries and the New Public Management movement prevalent in international discourse and practice. Drawing on private sector management concepts and tools, NPM is based on a ‘management cycle’ with a typical sequence of goal setting, implementation and evaluation. While this shows a marked conceptual kinship with the previous ‘policy cycle’, a profound difference is its constitutive and strategic ties to the ongoing activities of the operational unit concerned. Whereas the cost-efficiency–related evaluation of the ‘second wave’ was still largely conducted as external evaluation and was mainly meant to check and reduce (expansive and expensive) welfare-state policies, the evaluative activities and tools mandated by and following from the ‘management cycle’ are, first of all, of an internal nature: revolving around agency-based performance management, self-evaluative procedures and reporting, thus forming an integral part of the ‘public management package’ (see Furubo and Sandahl 2002, pp. 19 ff.). Thus, the ‘third wave’ is characterised by internal evaluative institutions and tools taking centre stage.

III. The many facets of public-sector reforms

The preceding summary of the three phases of public-sector reforms and evaluation during the past 30 years has hinted at great variation in the conceptual and institutional inventory of each phase.

The ‘planning period’ of the 1960s and 1970s engendered a broad spectrum of reform options which addressed the reorganisation of governmental and ministerial structures, decentralisation and deconcentration of political and administrative functions and territorial reforms, as well as the introduction of policy evaluation as an instrument of policy making.

In the ‘retrenchment period’ of the mid-1970s and 1980s, institutional changes were achieved through deregulation and the privatisation of public assets, while evaluation turned to cost-reducing procedures such as cost-benefit analyses and task scrutinies.

Finally, in the current period, NPM-guided institutional reforms, such as downsizing, agencification, contracting, outsourcing and performance management have been on the rise, along with concomitant evaluative procedures (performance monitoring and measurement, controlling, etc.).

From country to country—and even within each country—the mix of reform concepts and components being considered or implemented may vary greatly. For one thing, NPM is far from being a well-defined and consistent body of concepts. Instead, it is a bundle of different (and sometimes contradictory) concepts (see, for example, Aucoin 1990; Christensen and Lægreid 2001, p. 19). Picking and (eclectically) selecting from what has been somewhat ironically called a ‘shopping basket’ (Pollitt 1995), the

varied concepts and elements of NPM strategies and measures have been portioned and 'packaged' quite differently in different national, regional and local contexts.

Second, in many situations reform concepts and components which stem from previous reform periods may have persisted and may lend themselves to amalgamation with NPM-specific elements. Furthermore, the current modernisation thrust may open a window of opportunity for implementing or reviving previous reform concepts (such as decentralisation of political and administrative responsibilities). For the sake of analytical differentiation it seems advisable to make a distinction between traditional reform concepts and elements (particularly those of the 'planning period') and NPM concepts (in the narrow sense).

IV. The many faces and variants of evaluation

Departing from a broad understanding of the evaluation function

In order to capture the broad scope of pertinent analytical tools which have been put in place since the early 1960s for an evaluative purpose, on the one hand, a broad definition seems advisable. On the other hand, lest such a definition become a catchall (and thus ‘catch nothing’) concept, some delineation is, of course, needed.

At this point we depart from a broad understanding of evaluation as an analytical procedure and tool meant to obtain all information pertinent to the assessment of the performance, both process and result, of a policy program or measure. To be sure, a bewildering array of concepts and terms has made its appearance in this field, especially given the recent ‘third wave’ development of new vocabulary (such as management audit, policy audit and performance monitoring). In light of a definition which focuses on the function of evaluation and, thus, looks beneath the ‘surface’ of varied terminology, it becomes apparent that the different terms ‘cover more or less the same grounds’ (Bemelmans-Videc 2002, p. 94). Thus, analytical procedures which have come to be called ‘performance audit’ will be included in our definition, except, however, for ‘financial audit’ which checks the compliance of public spending with budgetary provisions and would not be included in evaluation (see Sandahl 1992, p. 115, Barzelay 1997: 235 ff. for a detailed discussion and references).

In the next sections further definitional distinctions and differentiations of ‘evaluation’ will be submitted.

Evaluation of public-sector reform policies and measures versus evaluation of 'substantive' policies

Substantive policies have been called, as it were, the 'normal' policies (such as social policy, employment policy and youth policy housing policy) which essentially target the socio-economic situation and development in the policy environment.

By contrast, public-sector reform policies are, by definition, directed at remoulding the political and administrative structures. Thus, one may speak in an institutionalist parlance of institution policy² or, in a policy science diction, of polity policy³ or of meta-policy making.⁴

Whereas 'substantive policies' can, simply stated, be seen as aiming directly at attaining their 'substantive' goal and policy 'output' (say, the reduction of unemployment or the improvement of the environment), public-sector reform policies have a more complicated 'architecture'.

As a first step they aim at effecting changes in political and administrative institutions as the immediate and 'closest' target of their intervention.

In the further sequence of goals, the institutional changes once effected are, in turn, intended to bring about further (and 'ultimate') results, be it that the operational process ('performance') of public administration or that the (final) 'output' and 'product' of the administrative operation is improved.

This sequence of goals can be translated into a corresponding set of evaluation questions.

The evaluation question can address whether and how the intended institutional changes (such as the creation of agencies, the intra-administrative decentralisation of

responsibilities and resources or the installation of benchmarking) have been achieved (or implemented). Owing to this implementation focus, one might speak of implementation evaluation⁵ (in fact, this evaluation variant conceptually has much in common with implementation research in political science, of which Pressman and Wildavsky's 1974 study was a pacesetter).

The evaluation question then may target the operational performance and 'process improvement' (Pollitt and Bouckaert 2000, p. 115) resulting from a reform measure (such as the 'speeding up' of administrative activities or their accessibility to citizens). One might speak of performance evaluation.

Finally, evaluation may be mandated to find out whether the output and outcomes of administrative activities have been affected by the reform. This may be termed output evaluation, impact evaluation or result evaluation (see Bemelsman-Videc 2002, p. 93).

But the reach of the evaluation may go still further, including more 'remote' effects such as 'systemic' effects (Pollitt and Bouckaert 2000, pp. 120 ff.) or impacts on the 'broader political-democratic context' (Christensen and Lægreid 2001, p. 32).

Monitoring versus evaluation research

Under methodological auspices monitoring can be seen as an evaluative procedure which aims at (descriptively) identifying and/or measuring the effects of an ongoing activity without raising the question of causality. In fact, in the 'third wave' of evaluation monitoring has come to play a pivotal role as an internal indicator-based and result-oriented procedure and tool of information gathering and reporting.

By contrast, evaluation research can be understood as an analytical exercise which typically employs social-scientific methodology. It is usually commissioned to tackle evaluation questions and projects of a higher complexity, typically posed by the ‘causal question’, that is, as to whether the observed result or output can be causally related to the policy ‘intervention’ (program component, activity) concerned.

When dealing with the evaluation of public-sector reform policies and measures evaluation research confronts methodological problems that are even thornier than in policy evaluation in general (see Pollitt 1995 and Pollitt and Bouckaert, Chapter 2 in this volume).⁶ A few potential methodological problems are these:

- 1) goals and objectives that serve as a measuring rod are hard to identify, particularly because modernisation measures mostly come in bundles;
- 2) goals are hard to translate into operationalisable and measurable indicators;
- 3) good empirical data to ‘fill in’ the indicators are hard to get, and the more meaningful an indicator is, the more difficult it is to obtain viable data;
- 4) the more ‘remote’ (and, often, the more relevant) the goal dimension is, the harder it becomes to operationalise and to empirically substantiate it (for example, outcomes, ‘systemic’ effects [see Pollitt and Bouckaert 2000, pp.120 ff], or effects on the ‘broader political-democratic context’ [see Christensen and Lægreid 2001, p. 32]);
- 5) side effects and unintended consequences⁷ are hard to trace; and
- 6) methodologically robust research designs (quasi-experimental, ‘controlled’ time-series, etc.) are rarely applicable (*ceteris paribus*

conditions difficult, if not impossible, to establish; N too small; ‘before’ data not available for a ‘before/after’ design; etc.).

‘Normal’ (‘primary’) evaluation versus meta-evaluation (‘secondary’ evaluation)

Meta-evaluation is meant to analyse an already completed (‘primary’) evaluation using a kind of ‘secondary’ analysis. Two variants can be discerned.

First, the meta-evaluation may review the already completed (‘primary’) evaluation in terms of whether it was done using an appropriate methodological approach. One might speak of a ‘methodology-related’ meta-evaluation.

Second, the meta-evaluation may have to accumulate the substantive findings of the already completed (‘primary’) evaluation and synthesise the results. This could be called a ‘synthesising’ meta-evaluation.

Internal versus external evaluation

An internal (‘in house’ or agency-based) evaluation is one conducted by the operating unit itself in an exercise of ‘self-evaluation’. In fact, the internal self-evaluation operation is a key procedure and component of the entire monitoring and feedback system which is pivotal to NPM’s (internal) management and accounting system.

External evaluation is initiated, and either conducted or funded and ‘contracted out’, by an agency or actor outside of and different from the operating unit. This external unit may exist within the core executive government (for instance, by the Finance Minister or the Prime Minister’s Office); it may be another

political/constitutional actor (particularly parliament or a court of audit); or it may be an organisation expressly created for that external evaluation function (such as an ad hoc commission or task forces).

In-house evaluation versus ‘contractual research’

In order to cope with a (methodologically or otherwise) complex piece of evaluation in light of limited analytical resources and competence, the agency and institution that initiates an external evaluation (or the operating unit itself, in the case of a methodologically demanding internal evaluation) may prefer to ‘contract out’ the evaluation to a self-standing (ideally independent) semi-public non-profit or university-based research institute or commercial research unit (such as a consultancy firm). In such a case, the evaluation is carried out by the (commissioned) research unit as contractual research (see Wollmann 2002a); the (commissioning) agency finances and monitors the ‘contractual’ evaluation and ‘owns’ the results thereof.

In distinction from, and in contrast with evaluation research as ‘commissioned’ (contractual) research on public-sector reforms mention should be made, at this point, of academic research which, following the ‘intra-scientific’ selection of topics, concepts and methods and funded by (independent) foundations or university resources, studies public-sector reform in an implementation or evaluative perspective with what may be called an ‘applied basic research’ approach.

Ex-ante, ongoing and ex-post evaluation

Reference should briefly be also made to the ‘classical’ distinction between ex-ante, ongoing/interim and ex-post evaluation.

Ex-ante evaluation is meant to anticipate and pre-assess the (alternative) courses of policy implementation (‘implementation pre-assessment’) and policy results and consequences (for instance, environmental impact assessments).

Ongoing evaluation has the task of monitoring and checking the processes and (interim) results of policy programs and measures while the implementation and realisation thereof is still going on. As ‘formative’ evaluation, it is designed to monitor and feed process data and (interim) result data back to the policy makers and project managers while the measure and project still is in its developmental and ‘formative’ stage, that is, in a stage that still allows the correcting and re-orienting the policy measures. As NPM hinges conceptually and instrumentally on the strategic idea of institutionalising permanent internal processes of data monitoring and (feedback) reporting, ongoing evaluation forms a central component of the ‘new public management package’.

Ex-post evaluation constitutes the classical variant of (substantive) policy and program evaluation, particularly in the full-fledged evaluation research type.

(Rigorous) evaluation versus ‘best practice’ accounts

While (rigorous) evaluation aims at giving a comprehensive picture of ‘what has happened’ in the policy field and project under scrutiny, encompassing successful as

well as unsuccessful courses of events, the best practice approach tends to pick ‘success stories’ of reform policies and projects, with the analytical intention of identifying the factors that explain the ‘success’, and with the ‘applied’ (learning and ‘pedagogic’) purpose to foster ‘lesson drawing’ from such experience in the intranational as well as in the inter- and transnational contexts. On the one hand, such ‘good practice’ stories are fraught with the (conceptual and methodological) threat of ‘ecological fallacy’, that is, of a rash and misleading translation and transfer of (seemingly positive) strategies from one locality and one country to another. On the other hand, if done in a way which carefully heeds the specific contextuality and conditionality of such ‘good practice’ examples, analysing, ‘telling’ and diffusing such cases can be useful ‘fast track’ to evaluative knowledge and intra-national as well as trans-national learning (see Jann and Reichard, chapter 3 in this volume).

Quasi-evaluation: Evaluation as an interactive learning process

Vis-à-vis these manifold conceptual and methodological hurdles ‘full-fledged’ evaluation of public-sector reforms is bound to face (and also in light of the reluctance which policy-makers and top administrators often exhibit towards getting researchers from outside intimately involved in ‘in-depth’ evaluations), Thoenig proposes (in Chapter 11 of this volume) a type of ‘quasi-evaluation’ which would be less fraught with conceptual and methodological predicaments than a ‘full-fledged’ evaluation and more disposed toward focusing on, and restricting itself to, the information- and data-gathering and descriptive functions of evaluation rather than an explanatory one.

Thoenig perceives more than one advantage to the ‘quasi-evaluation’ approach. First,

such conceptually and methodologically ‘lean’ evaluation designs may find easier access and wider application in the otherwise in an evaluation territory otherwise fraught with hurdles (he caustically remarks that ‘there is no surer way of stifling evaluation at the outset than to confine it to the ghetto of methodology’ [see chapter 11 in this volume]). Second, a conceptually and methodologically pared-down variant of ‘quasi-evaluation’ may be conducive to more ‘trustful’ communication between the policy-maker and the evaluator and to promote a ‘gradual learning process that fosters an information culture’ (Chapter 11 in this volume).

1. Important exceptions are Pollitt and Bouckaert 2000 (which covers public-sector reform in 10 OECD countries, including the Netherlands, France, and Germany) and Furubo, Rist, and Sandahl's 2002 'atlas' of evaluation, which contains as many as 21 country reports.
2. On the distinction between substantive policy (substanzielle Politik) and institution policy (Institutionenpolitik) see Knoepfel and Bussmann 1997, p. 59, and Ritz 1999, p. 28.
3. This refers to the distinction made in policy science and policy studies between policy (as the contents of policy making), politics (as the process of policy making) and polity (as the institutional setting thereof).
4. This term was coined by Yezekel Dror.
5. See Christensen et al. in chapter 4 of this volume: 'Process evaluation tracks the extent to which program or practices were put in place as intended and monitor how implementation has progressed'.
6. For a penetrating discussion of the methodological issues of evaluation (research) at large, see Pawson and Tilley 1997.
7. See Jann and Reichard in Chapter 3 of this volume: 'No organizational change of even modest complexity will happen without the most common of all social phenomena: unintended and even counterintuitive processes and results'.

References

- Aucoin, Peter (1990), 'Administrative reform in public management. Paradigms, principles, paradoxes and pendulums', Governance, 3[2], 115-137.
- Barzelay, M. (1997), 'Central Audit Institutions and Performance Auditing: A Comparative Analysis of Organizational Strategies in the OECD, in Governance, vol. 10, no. 3, pp. 235-260
- Bemelmans-Videc, M.L. (2002), 'Evaluation in The Netherlands 1990-2000.Consolidation and Expansion', in Jan-Eric Furubo, Ray C. Rist and Rolf Sandahl (eds), International Atlas of Evaluation, New Brunswick and London: Transaction, pp. 115-128.
- Christensen, Tom and Per Lægreid (2001), 'A Transformative Perspective on Administrative Reforms', in Tom Christensen and Per Lægreid (eds), New Public Management, Aldershot: Ashgate, pp. 13-39.
- Derlien, Hans-Ulrich (1990), 'Genesis and Structure of Evaluation Efforts in Comparative Perspective', in Ray C Rist (ed.), Program Evaluation and the Management of Government, New Brunswick and London: Transaction, pp. 147-177.
- Furubo, Jan-Eric, Ray C. Rist and Rolf Sandahl(eds) (2002), International Atlas of Evaluation, New Brunswick and London: Transaction.
- Furubo, Jan-Eric and Rolf Sandahl (2002), 'A Diffusion-Perspective on Global Developments in Evaluation', in Jan-Eric Furubo, Ray C. Rist and Rolf Sandahl (eds), International Atlas of Evaluation, New Brunswick and London: Transaction, pp. 1-26.

- Hood, Christopher (1991), 'A public management for all seasons?', Public Administration, 69[Spring], 3-19.
- Knoepfel, Peter and Werner Bussmann (1997), 'Die öffentliche Politik als Evaluationsobjekt', in Werner Bussmann, Ulrich Klöti and Peter Knöpfel (eds), Einführung in die Politikevaluation, Basel: Helbing & Lichtenhan, pp. 58-77.
- Levine, Robert A. (1981), 'Program Evaluation and Policy Analysis in Western Nations: An Overview', in Robert A. Levine, Marian A. Solomon, Gerd-Michael Hellstern, and Hellmut Wollmann (eds), Evaluation Research and Practice: Comparative and International Perspectives, Beverly Hills and London: Sage, pp. 12-27.
- Pawson, Ray and Nick Tilley (1997), Realistic Evaluation, London: Sage.
- Pollitt, Christopher (1995), 'Justification by works or by faith? Evaluating the New Public Management', Evaluation, 1[2 (October)], 133-154.
- Pollitt, Christopher and Geert Bouckaert (2000), Public Management Reform, Oxford: Oxford University Press.
- Pressman, Jeffrey and Aaron Wildavsky (1974), Implementation (1984 3rd ed.), Berkeley: University of California Press.
- Rist, Ray C. (ed.) (1990), Program Evaluation and the Management of Government, New Brunswick and London: Transaction.
- Ritz, Adrian (1999), Die Evaluation von New Public Management, Bern: IOP-Verlag.
- Sandahl, Rolf (1992), 'Evaluation at the Swedish National Audit Bureau', in J. Mayne et al. (eds), Advancing Public Policy Evaluation, Amsterdam: Elsevier, pp. 115-121.

- Sanderson, Ian (2000), 'Evaluation in Complex Policy Systems', Evaluation, 6[4], 433-454.
- Vedung, Evert (1997), Public Policy and Program Evaluation, New Brunswick: Transaction.
- Wagner, Peter and Hellmut Wollmann (1986), 'Fluctuations in the development of evaluation research: Do regime shifts matter?', International Social Science Journal, 108, 205-218.
- Wollmann, Hellmut (ed.) (2001), 'Evaluating Public Sector Reforms', special issue of Revista Internacional de Estudios Politicos, 127-143.
- Wollmann, Hellmut (2002a), 'Contractual Research and Policy Knowledge', International Encyclopedia of the Social and Behavioral Sciences, 5, 11574-11578.
- Wollmann, Hellmut (2002b), 'Verwaltungspolitik und Evaluierung: Ansätze, Phasen und Beispiele im Ausland und in Deutschland, Evaluation und New Public Management', Zeitschrift für Evaluation, 1, 75-101.