Evaluation and evaluation research

Hellmut Wollmann

to be published in: Fischer, Frank/ Miller, Gerald/ Sidney, Mara (eds.), Handbook of Poublic Policy Analysis. Theory, Politics and Methods, New York: Marcel Dekker Publisher (forthcoming)

Definitions, Concepts, and Types of Evaluation

Evaluation in the field of public policy may be defined, in very general terms, as an analytical tool and procedure meant to do two things. First, evaluation research, as an analytical tool, involves investigating a policy program to obtain all information pertinent to the assessment of its performance, both process and result; second, evaluation as a phase of the policy cycle more generally refers to the reporting of such information back to the into the policy-making process (see Wollmann, 2003b: 4).

Yet, 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 would be included in our definition, except, however, 'financial audit' which checks the compliance of public spending with budgetary provisions and would not counted as evaluation (see Sandahl,1992: 115).

Types of Evaluation: Functions and Timing

In terms of the different temporal and functional linkages with the *policy cycle* often he following distinctions are made (see Wollmann, 2003b):

Ex ante evaluation, preceding decision-making, is meant to (hypothetically) anticipate and pre-assess the effects and consequences of planned or defined policies and actions in order to "feed" the information into the upcoming or ongoing decision-making process.. If undertaken on alternative courses of policies and actions, ex ante evaluation is an instrument of making the choice between alternative policy options (ideally) analytically more transparent, more

foreseeable and politically more debatable. *Implementation pre-assessment* is meant to analytically anticipate the course of policy implementation in focusing on its process, as well as e*nvironmental impact assessment,* designed for anticipating or predicting the consequences which envisaged policies and measure may have on the environment.

Ongoing evaluation has the task of identifying the (interim) effects and results of policy programs and measures while, in the policy cycle, the implementation and realisation thereof is still under way. The essential function of ongoing evaluation is to feed relevant information back into the implementation process at a point and stage when pertinent information can be used in order to adjust, correct or redirect the implementation process or even underlying key policy decisions. In a nearly synonymous usage, some speak of *accompanying* evaluation running parallel to the policy implementation process. Within ongoing or accompanying evaluation one can discern between a primarily analytical modality which remains detached and distanced from the implementation process in order to ascertain objectivity. Further, the term *interventionist* (accompanying) evaluation has been applied when, besides the analytical mandate, the evaluators are also expected, if not obliged to actively intervene in the implementation process in order to rectify shortcomings and flaws in the implementation process jeopardising the attainment of the pre-set policy goals. In such an interventionist orientation "accompanying" evaluation would approximate the concept of *action research*.

Finally, monitoring can be seen as an (on-going) evaluative procedure which aims at (descriptively) identifying and, with the help of appropriate, if possible operationalised, indicators, at measuring the effects of ongoing activities. In the most recent upsurge of *performance indicators* (PIs) in the concepts of New Public Management, indicator-based monitoring has gained great importance.

Ex-post evaluation constitutes the classical variant of evaluation to assess the goal attainment and effects of policies and measures, once they have been completed As such *summative* evaluation (Scriven, 1972) has been directed at policy *programs* (as a policy form that combines the setting of policy goals with the financial, organisational, personnel etc. resources meant to achieve these goals which was typical of the conduct of reform policies in the US but also in European countries), ex post policy evaluation has often been identified with *program evaluation* (see Rist ed., 1990). Characteristically policy (or program) evaluation has been given primarily two tasks.

First, it was meant to produce an assessment about the degree to which the intended policy goals have achieved (goal attainment). The *conceptual problems* following from this task revolve around the conceptualising the appropriate, if possible measurable, indicators in order to make such assessments of goal attainment. But, besides identifying the *intended* consequences, the assessment of the effects of policies and programs came to pertain also to the *non-intended* consequences.

Second, the evaluation of policies and programs was also expected and mandated to answer the (*causal)* question as to whether the observed effects and changes have be really (causally) related to the policy or program in question.

Meta-evaluation is meant to analyse an already completed ('primary') evaluation using a kind of 'secondary' analysis. Two variants may be discerned. First, the meta-evaluation may review the already completed piece of (primary) evaluation as to whether it is up to methodological criteria and standards. One might speak of *methodology-focused* meta-evaluation. Second, the meta-evaluation may be meant to accumulate the substantive findings of the already completed ('primary') evaluation and synthesise the results. This might be called a *finding-focused* meta-evaluation.

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 up and 'tell 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, 'story-telling' and diffusing such cases can provide a useful fast track to evaluative knowledge and intra-national as well as trans-national learning

Vis-à-vis these manifold conceptual and methodological hurdles fully fledged evaluation of public-sector reforms is bound to face a type of *quasi-evaluation* has been proposed (see Thoenig, 2003) that would be less fraught with conceptual and methodological predicaments than a 'full-scale' 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. A major assets may be 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' (Thoenig, 2003).

Finally, an evaluability assessment can be undertaken. This happens before an evaluation, be it of the ex post, but also of the ex-ante and on-going type. It is used to find out in advance whether and which approach and variant of evaluation should be turned to on the basis of the criteria of technical feasiblity, economic viability and/or of practical merits.

As 'classical' evaluation is, first of all, directed at (ex post) assessing the attainment or non-attainment of the policy and program goals or at (ex ante) estimating the attainability of

goals, it deals essentially with the *effectiveness* of policies and measures which amount of resources has been put up in order to reach that goal. This stands in contrast to a *cost-benefit-analysis* which compares the outcomes to the resources devoted to achieve them.

.

Types of Evaluation: Internal and External

For one, evaluation may be conducted as *internal* evaluation. Such evaluation is carried out *in-house* by the operating agency itself. In this case, it takes place as *self-evaluation*. In fact, one might argue that informal and unsystematic modes of self-evaluation have been practiced ever since in the (*Weberian*) bureaucracy model (hierarchical) oversight has taken place based on forms of regular internal reporting. But evaluation research involves more formal approaches. The have become key components of various theories of public administration. In recent years New Public Management has emphasized the concept of monitoring and controlling based on evaluation performance indicators. They play, for example, a pivotal role in operating systems of comprehensive internal cost-achievement accounting (see Wollmann, 2003b).

External evaluation, by contrast, is initiated or funded by outside sources (contracted out by an agency or actor outside of the operating administrative unit). Such an external locus of the evaluation function may be put in place by institutions and actors that, outside and beyond administration, may have a political or structural interest employing evaluation as a means to oversee the implementation of policies by administration. *Parliaments* have shown to be the natural candidates for initiating and carrying out the evaluation of policies and programs inaugurated by them. In a similar vein, courts of audits have come to use evaluation as an additional analytical avenue to shedding light on the effectiveness and efficiency of administrative operations.

But also other actors within the 'core' government, such as the Prime Minister's Office or the Finance Ministry, may turn to evaluation as an instrument to oversee the operations of sectoral ministries and agencies. Finally, mention should be made of *ad hoc* bodies and commissions (of the Royal Commission or enquiry commission type) which, being mandated to scrutinise and to come out with recommendations on complex issues and entire policy fields, including the policy implementation by government and ministries, may employ evaluation as an important fact-finding tool.

The more complex the policies and programs under consideration are, and the more demanding the conceptual and methodological problems of carrying out such evaluations become, the less the institutions, initiating and conducting the evaluation, are capable to carry out such conceptually and methodologically complicated and sophisticated analyses themselves. In view of such complexities, evaluation research is ideally based on the application of social science methodology and expertise. Thus, in lack of adequately trained personnel and of time the political, administrative and the other institutions often turn to outside (social science) research institutes and research enterprises in commissioning them to carry out the evaluation work on a *contractual* basis (see Wollmann, 2002). In fact, the development of evaluation, since the mid 1960s, has been accompanied by the (at times rampant) expansion of a 'contractual money market' which, fed by the resources of ministries, parliament, adhoc commissions etc, has turned evaluation research virtually into a "new industry of considerable proportion" (Freeman/ Solomon, 1981: 13), revolving around *contractual research*, and has deeply remolded the traditional research landscape in a momentous shift from *academic to entrepreneurial* research (see Freeman/Solomon, 1981: 16), a topic to which we return.

The 'three waves' of evaluation

Roughly three phases ('waves') can be distinguished in the development of evaluation over the past 40 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 one setting in since the 1990.

During the 1960s and 1970s the advent of the advanced welfare state was accompanied by the concept of enhancing the ability of the state for proactive policy making through the modernisation of its political and administrative structures in the pursuit of which the institutionalisation and employment of planning, information and evaluation capacities was seen instrumental. The concept of a *policy cycle* revolved, as already mentioned, around the triad of policy formation, implementation and termination, whereby evaluation was deemed crucial as a cybernetic loop in gathering and feeding back policy-relevant information The underlying *scientific logic* (Wittrock, Wagner, Wollmann, 1991: 615) and vision of a science-driven policy model was epitomised by *Donald Campbell´*s famous call for an *experimenting society* ('reforms as experiments', Campbell, 1969) .

In the *United States* the rise of evaluation came with the inauguration of federal social action programmes ('War on Poverty') in the mid-1960s under President Johnson with evaluation almost routinely mandated by the pertinent reform legislation, turning policy and program evaluation virtually into a veritable growth industry. Large-scale social experimentation with accompanying major evaluation followed suit. In Europe, Sweden, Germany and the U.K. became the front-runners of this first wave of evaluation (see Levine, 1981, Wagner, Wollmann, 1986, Derlien, 1990) whereby in Germany social experimentation (*experimentelle Politik)* was undertaken on a scale unparalleled outside the United States (see Wagner, Wollmann, 1991: 74).

Reflecting the reformist consensus which at that time was widely shared by reformist political and administrative actors as well as by the social scientists, involved through hitherto largely unknown forms of contractual research and policy consultancy, the then conducted evaluation projects normatively agreed with and supported the reformist policies under scrutiny

and were, hence, meant to **i**mprove policy results and to maximise output effectiveness. (Wittrock, Wagner, Wollmann, 1991: 52)

The heydays of the interventionist welfare state policies proved to be short-lived, when, following the first oil price rise of 1973, the world economy slid into a deepening recession and the national budgets ran into a worsening financial squeeze which brought most of the cost-intensive reform policies to a grinding halt. This lead to the second wave**.** As policy making came to be dictated by the calls for budgetary retrenchment and cost-saving, the mandate of policy evaluation got accordingly redefined with the aim to reducing the costs of policies and programs, if not to phase them out (see Wagner, Wollmann, 1986, Derlien, 1990). In this *second wave* of evaluation focusing on the cost-efficiency of policies and programs evaluation saw a significant expansion in other countries, for instance, in the Netherlands (see Leeuw 2004, 60).

A *third wave* can be identified since the 1990s.under the influence of sundry currents. For one, the concepts and imperatives of *New Public Management* (see Pollitt, Bouckaert, 2004) have come to dominate the international modernisation discourse and, in one or the other variant, the public sector reform in many countries (see Wollmann, 2003c). Hereby *internal evaluation* (through the build-up and employment of indicator-based controlling and cost-achievement-accounting etc) formed integral part of the 'public management package' (see Furubo, Sandahl, 2002, pp. 19 ff.) and gave new momentum to evaluative procedures (see Wollmann, 2003b.). Moreover, in a number of policy fields evaluation has gained salience in laying bare the existing policy shortcomings and in identifying the potential for reforms and improvements. The great attention (and excitement) raised recently by the (OECD-wide) *PISA study* as a major international evaluation exercise on the national educational systems has highlighted and, no doubt, propelled the role and potential of evaluation as an instrument of policy making. Thirdly, mention should be made that, within the European Union, evaluation has been given a major push when the European Commission decided to have the huge spending of the European Structural Fund systematically evaluated

(see Leeuw, 2004: 69 ff.). As EU's structural funds are now being evaluated, within their five year program cycle, in an almost text book-like fashion (with an evaluation cycle running from ex ante- through ongoing to ex post- evaluation), the evaluation of EU policies and programs has significantly influenced and pushed ahead the development of evaluation at large. In some countries, for instance in Italy (see Stame, 2002, Lippi, 2003)..the mandate to evaluate EU programs was, as it were, the cradle of the country's evaluation research which had hardly existed before.

In an international comparative perspective, then, at the beginning of the new millennium, policy evaluation has been introduced and installed in many countries as a widely accepted and employed instrument of gaining (and of feeding back) policy-relevant information. This has been impressively analysed and documented in a recent study[1] based on country reports on 22 countries and on a sophisticated set of criteria (see Furubo, Rist, Sandahl eds., 2002, with a summarising and synthesising piece Furubo, Sandahl, 2002). While the US are still holding the lead in the *evaluation culture* (Rist, Pakiolas, 2002: 230 ff.), the upper six ranks among European countries are taken by Sweden, the Netherlands, UK, Germany, Denmark and Finland (see Furubo, Sandahl, 2002, Leeuw, 2004: 63).

Methodological Issues of Evaluation

The main conceptual and methodological tasks which evaluation research is faced with are (1) to conceptualise the observable real world changes in terms of intended (or non-intended) consequences which policy evaluation is meant to identify and to assess (as, methodologically speaking, dependent variable), and (2) to find out whether and how the observed changes are ('causally') linked to the policy and measure under consideration (as independent variable).

In coping with these key questions, evaluation research is seen to be an integral part of social science research at large; it includes, as such, most of social science's conceptual and

methodological issues and controversies. In fact, it seems that the methodological debates which have occurred in the social science community at large (for instance in the strife between the quantitative and the qualitative schools of thought) appears to have had one of its most pronounced (and at times fiercest) battle-ground in the evaluation research community.

Two phases can be discerned in this controversy. The first, dating from the 1960s to the early 1980s, has been characterised by the dominance of the neopositivist-nomological science model (with an ensuing preponderance of the quantitative and –quasi-experimental methods). The second and more recent period has resulted from advances in the constructivist, interpretive approach (with a corresponding preference for qualitative heuristic methods).

Accordingly, from the neopositivist perspective, evaluation has been characterised by two premises. The first is the assumption that in order to validly assess whether and to which degree the policy goals (as intended consequences) have been attained by observable real world changes, it is necessary to identify in advance what the political intentions and goals of the program are. In this view, the intention of the one relevant institution or actor stands in the fore.

]Second, in order to identify *causal* relations between the observed changes and the policy/program under consideration, valid statements could be gained only through the positivist application of quantitative, (quasi-) experimental research designs. (Campbell, Stanley, 1963). Yet, notwithstanding the long dominance of this research paradigm, the problem of translating these premises into evaluation practice were obvious to many observers. For example, in identifying the relevant objectives serious issues arise (see Wollmann, 2003b: 6): (1) goals and objectives that serve as a measuring rod are hard to identify, as they often come as ‘bundles’ of goals that 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; (3) the more 'remote' (and, often, the more relevant) the goal dimension is, the harder it becomes to operationalise and to empirically substantiate it; (4) side effects and unintended consequences are hard to trace.

Moreover, methodologically robust research designs (quasi-experimental, controlled time-series, etc.) are often not applicable, at least not in a methodologically satisfying manner (Weiss, Rein, 1970) Often enough the *.ceteris paribus* conditions (on which the application of quasi-experimental designs hinges) are difficult, if not impossible, to establish. While the application of quantitative methods is premised on the methodological 'rule of thumb' of requiring 'many cases (large N), but few variables', in the real world of research the constellation is often quite opposite with 'few cases (small N), but many (possibly relevant) variables'. These problems tend to rule out the employment of quantitative methods and, instead, suggest to resort to qualitative approaches and methods. And finally, the application of *time series* methods (*before/after* design) has often narrow limits, as the *before* data are often not available nor procurable.

In the second phase, the long dominant research paradigm has come under criticism on two interrelated scores. For one, the standard assumption that evaluation should seek its frame of reference first of all in the policy intention of the relevant political institution (s) or actor(s) has been shaken--if not shattered--by the advances of the *constructivist-interpretive* school of thought (Mertens, 2004: 42 ff.). Its advocates questioning on epistemological grounds, the possibility of validly ascertaining one relevant intention or goals and call instead for identifying a plurality of perspectives, interests and values. For instance, Stufflebeam (1983) has been influential in advancing a concept of evaluation called the *CIPP model* in which C= context, I = input, P = process, P = product). Among the four components, the context element (focusing on questions like: What are the program's goals? Do they reflect the needs of the participants?) is meant to direct evaluator's attention, from the outset, to the needs (and

interests) of the participants of the program under consideration (and its underlying normative implications). This general line of argument has been expressed in different formulations, such as *responsive*, *participatory* or *stakeholder* evaluation. Methodologically the constructivist debate has gone hand-in-hand with (re-) gaining ground for qualitative-hermeneutic methods in evaluation (Mertens, 2004: 47). Guba and Lincoln (1989) have labelled this development 'fourth generation evaluation'.

While the battle-lines between the schools and camps of thought were fairly sharply drawn some twenty years ago, they have softened up in the meantime. On the one hand, the epistemological, conceptual and methodological insights generated in the constructivist debate are accepted and taken seriously, the mandate in evaluation to come as close as possible to 'objectivity' still remains a major goal. The concept of a 'realistic evaluation' as formulated by Pawson, Tilley (1997) lends itself to serve that purpose. Furthermore, it is widely agreed that there is no king's road in the methodological design of evaluation research; instead, one should acknowledge a pluralism of methods. The selection and combination of the specific set and mix of methods depends on the evaluative question to be answered, as well as the time frame and financial and personnel resources available.

Evaluation Research : Between Basic, Applied, and Contractual Research

The emergence and expansion of evaluation research since the mid-1960s has had a significant impact on the social science research landscape and community. Originally the social science research arena was dominated by *academic (basic)* research primarily located at the universities and funded by independent research funding agencies. Even when it took an *applied policy* orientation, social science research remained essentially committed to the academic/basic formula. By contrast, evaluation research, insofar as it is undertaken as "contractual research", commissioned and financed by a political or administrative institution, involves a shift from "academic to enterpreneurial settings" (Freeman, Solomon, 1981).

Academic social science research, typically university-based, has been premised on four imperatives. The first has been a commitment to *seek the truth* as the pivotal aim and criteria of scientific research. The second relates to intra-scientific autonomy in the selection of the subject-matter and the methods of its research. The third has been independent funding, be it from university sources or through peer review-based funding by research foundations such as the National Science Foundation. And the final component has been the testing of the quality of the research findings to an open scientific debate and peer-review.

While applied social science still holds on to the independence and autonomy of social science research, *contractual research,* which now constitutes a main vehicle of evaluation research, hinges on a quite different formula. It is characterised by a commissioner/producer or consumer/contractor principle: "the consumer says what he wants, the contractor does it (if he can), and the consumer pays" (to quote Lord Rothschild's dictum, see Wittrock, Wagner, Wollmann,. 1991: 47). Hence, the request for proposal (RFP) through which the commissioning agency addresses the would-be contractors (in public bidding, selective bidding or directly), generally defines and specifies the questions to be answered and the time frame made available. In the project proposal the would-be contractor explains his research plan within the parameters set by the 'customer' and makes his financial offer which is usually calculated on a personnel costs plus overheads formula.

Thus, when commissioned and funded by government, evaluation research confronts three crucial challenges related to the subject-matter, the leading questions, and the methods of its research. In contract research, unlike traditional evaluation research, these considerations are set by the agency commissioning the evaluation. Also, by providing the funding, the agency also jeopardises the autonomy of the researchers ('who pays the piper, calls the tune'). And finally, the findings of commissioned research are often held in secret, or at least are not published, thus bypassing an open public and peer debate. So, contractual research is exposed and may be vulnerable to an *epistemic drift* and to a *colonisation process* in which the evaluators may

induced to adopt the perspective and conceptual framework of the political and administrative institutions and actors they are commissioned to evaluate (Elzinga, 1983: 89).

In the face of the challenges to the intellectual integrity and honesty of contractual research, initiatives have taken by professional evaluators to formulate standards that could guide them in their contractual work, in particular in their negotiations with their 'clients' (Rossi, Freeman, Lipsey, 1999: 425 ff). Reference can be made here, for example, to *Guiding principles of Evaluation,* adopted in 1995 by the *American Evaluation Association* in 1995. Among its five principles the maxims of integrity and honesty of research are writ large (Rossi, Freeman, Lipsey, 1999: 427 ff.; and Mertens, 2004: 50 ff).

Professionalization

In the meantime evaluation has, in many countries, become an activity and occupation of a self-standing group and community of specialised researchers and analysts whose increasing professionalisation is seen in the formation of professional associations, the appearance of professional publications and in the arrival of evaluation as a subject matter in university and vocational training.

As to the foundation of professional associations, a leading and exemplary role was assumed by the *American Evaluation Society* which was formed in 1986 through the merger of two smaller evaluation associations, *Evaluation Network* and the *Evaluation Research Society*. As of 2003, AEA had more than 3.000 members (see Mertens, 2004: 50). An important product was the formulation of the aforementioned professional code of ethics laid down in the Guiding Principles for Evaluators adopted by the AES in 1995. In Europe, the *European Evaluation Society* was founded in 1987 and the establishment of national evaluation societies followed suit, with the *UK Evaluation Society* being the first[2] (see Leeuw, 2004: 64 f.).. In the meantime most of them have also elaborated and adopted professional

codes of ethics which expresses the intention and resolve to consolidate and ensure evaluation as a new occupation and profession.

Another important indicator of the professional institutionalisation of the evaluation is the extent to which evaluation has become the topic of a mushrooming publication market. This, not least, includes the publication of professional journals, often in close relation to the respective national association. Thus, the American Evaluation Association has two publications: The *American Journal of Evaluation* and the *New Directions for Evaluation* monograph series (see Mertens, 2004: 52). In Europe, the journal *Evaluation* is published, associated with the European Evaluation Society. Furthermore, a number of national evaluation journals (in the respective national language) have been started in a number of European countries. All of these serve as useful sources of information on the topic of evaluation research.

Notes

[1] For example, see the 'New Jersey Negative Income Tax experiment,' which involved $8 million for research spending (Rossi and Lyall 1978).

[2] For earlier useful overviews, see Levine et al. ed. 1981, Levine 1981. Wagner and Wollmann 1986, Rist ed. 1990, Derlien 1990, Mayne et al ed.. 1992

[3] European Evaluation Society **www.europeanevaluation.org**

Associazione Italiana de Valuatazione www.valutazione.it

Deutsche Gesellschaft für Evaluation www.degeval.de

Finnish Evaluation Societ e-mail: petri.virtanen@vm.vn.fi

Schweizerische Evaluationsgesellschaft www.seval.ch

Société Française de l'Evaluation www.sfe.asso.fr

Société Wallonne de l'Evaluation et de la Prospective www.prospeval.org

UK Evaluation Society www.evaluation.org.uk

References

Campbell, D. T. (1969), Reforms as Experiments, In American Psychologist, 1969, pp. 409-430

Campbell, D. T., Stanley, Y. (1963), Experimental and quasi-experimental evaluations in social research. Rand McNally, Chicago

Bemelmans-Videc, M.L. (2002), 'Evaluation in The Netherlands 1990-2000.Consolidation and Expansion', in J.-E. Furubo, R. C. Rist and R. Sandahl (eds), International Atlas of Evaluation, New Brunswick and London: Transaction, pp. 115-128.

Derlien, H.-U. (1990), 'Genesis and Structure of Evaluation Efforts in Comparative Perspective', in R.C. Rist (ed.), Program Evaluation and the Management of Government, New Brunswick and London: Transaction, pp. 147-177.

Freeman, H., Solomon, M.A. (1981), The Next Decade of Evaluation Research, In R.A. Levine, M.A. Solomon, G.-M. Hellstern and H. Wollmann, (eds.), Evaluation Research and Practive. Comparative and international perspectives, Beverly Hills: London: Sage, pp. 12-26

Furubo, J. , Rist, R. C, Sandahl, Rolf, eds. (2002), International Atlas of Evaluation, New Brunswick and London: Transaction.

Furubo, J.-E. , Sandahl R. (2002), 'A Diffusion-Perspective on Global Developments in Evaluation', in J.-E. Furubo, R. C. Rist and R. Sandahl (eds), International Atlas of Evaluation, New Brunswick and London: Transaction, pp. 1-26.

Elzinga, A. 1985, Research Bureaucracy and the Drift of Epistemic Criteria, in: B. Wittrock and A. Elzinga (eds.), The University Research System, Stockholm: Almqvist and Wiksell, pp. 191-220.

Guba, Y., Lincoln E. 1989, Fourth Generation Evaluation, London: Sage.

Lasswell, H. D. 1951. The Policy Orientation. In: D. Lerner and H.D. Lasswell (eds.), The Policy Sciences,. Stanford University Press. pp. 3-15

Leeuw, F. L. 2004, Evaluation in Europe, In R. Stockmann (ed.), Evaluationsforschung, 2d. ed., Wiesbaden: VS Verlag, pp. 61-83

Levine, R..A., Solomon, M. A., Hellstern, G., Wollmann, H, eds. (1981), Evaluation Research and Practive. Comparative and international perspectives, Beverly Hills/London: Sage

Levine, R.A. (1981), Program Evaluation and Policy Analysis in Western Nations: An Overview, in: R. Levine, M.A. Solomon,, G.-M. Hellstern and H. Wollmann (eds.), Evaluation Research and Practive. Comparative and international perspectives, Beverly Hills/London: Sage, pp. 12-27

Lippi, A. (2003), 'As a voluntary choice or as a legal obligation?' Assessing New Public Management policy in Italy, in: H. Wollmann (ed.), Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar, pp. 140-169

Mayne, J.L./ Bemelmans-Videc, M.L./ Hudson, J./ Conner, R., eds. (1992), Advancing Public Policy Evaluation, Amsterdam: North-Holland

Mertens, D.M. (2004), Institutionalising Evaluation in the United States of America, in: R. Stockmann (ed.), Evaluationsforschung, 2d. ed., Wiesbaden: VS Verlag, pp. 45-60

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

Pollitt, C. (1995), 'Justification by works or by faith? Evaluating the New Public Management', Evaluation, 1[2 (October)], 133-154.

Pollitt, C. and G. Bouckaert (2003), Evaluating public management reforms. An international perspective, in: H. Wollmann (ed.) , Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar, pp. 12-35

Pollitt, C. and G. Bouckaert (2004), Public Management Reform, 2nd ed., Oxford: Oxford University Press

Rossi, P., H.Freeman and M.W. Lipsey (1999), Evaluation. A systematic approach, 6th edition, Thousand Oaks etc.: Sage.

Rist, R., ed. (1990), Program Evaluation and the Management of Government, New Brunswick/London: Transaction.

Rist,R. and K. Paliokas (2002), The rise and fall (and rise again?) of the evaluation function in the US government, in: J.-E. Furubo, R.C. Rist and R., eds., International Atlas of Evaluation, New Brunswick and London: Transaction. pp. 225-245

Sandahl, R. (2002), Evaluation at the Swedish national audit bureau, in: J.L. Mayne, M.L. Bemelmans-Videc, J.Hudson and R. Conner, eds., Advancing Public Policy Evaluation, Amsterdam: North-Holland, pp. 115-121

Scriven, M. (1972), The Methodology of Evaluaiton, in: C.H. Weiss, ed., Evaluating Action Programs, Boston 1972, pp. 123 ff.

Stufflebeam, D.L. (1983), The CIPP model for program evaluation, in: G.F. Madaus, M. Scriven and D.L. Stufflebeam, eds., Evaluation Models, Boston: Kluwer-Nijhoff, pp. 117-142

Stame, N. (2003), Evaluation in Italy. An inverted sequence from performange managent toprogram evaluaton?, in: J.-E. Furubo, R. C. Rist and R. Sandahl, eds., International Atlas of Evaluation, New Brunswick and London: Transaction. pp. 273-290

Thoenig, J.-C. (2003), Learning from Evaluation Practice: The Case of Public-Sector Reform, In H. Wollmann, ed., Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar, pp. 209-230

Vedung, E. (1997), Public Policy and Program Evaluation, New Brunwick: Transaction.

Wagner, P. and Wollmann, H. (1986), 'Fluctuations in the development of evaluation research: Do regime shifts matter?', International Social Science Journal, 108, 205-218.

Wagner, P. and Wollmann, H. (1991). Beyond Serving State and Bureaucracy: Problem-oriented Social Science in (West) Germany. Knowledge and Policy, Vol. 4, Nos. 12, pp. 46-88.

Weiss, R.S and Rein, M. (1970), The Evaluation of broad-aim programs. Experimental Design, its difficulties and an alternative, Administrative Science Quarterly, 1970, pp.97 ff.

Wittrock, B., Wagner, P, Wollmann, H. (1991): Social science and the modern state, In P. Wagner, C.H.Weiss and H. Wollmann, eds., Social Sciences and Modern State, Cambridge: Cambridge University Press, pp. 28-85.

Wollmann, H. (2002), Contractual Research and Policy Knowledge, In International Encyclopedia of Social and Behavioral Sciences, vol. 5, pp. 11574 – 11578

Wollmann, H. ed. (2003a), Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar

Wollmann, H. (2003b), Evaluation in Public-Sector Reform. Towards a ‚third wave' of evaluation, In H. Wollmann ed., Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar, pp. 1-11

Wollmann, H. (2003c), Evaluation in Public-Sector Reform. Trends, Potentials and Limits in International Perspective, In H. Wollmann (ed.), Evaluation in Public-Sector Reform, Cheltenham: Edward Elgar, pp. 231-258

Wollmann, H. (2005), Applied social science : Development, state of the art, consequences, In UNESCO (ed.), History of Humanity, vol. VII, chapter 21, Routledge (forthcoming)