Evaluation Systems What Are They and Why Study Them?

Article *in* Evaluation · April 2008

DOI: 10.1177/1356389007087537

2 authors:

Frans Leeuw
Maastricht University
168 PUBLICATIONS **4,597** CITATIONS

SEE PROFILE

Jan-Eric Furubo
22 PUBLICATIONS **412** CITATIONS

SEE PROFILE

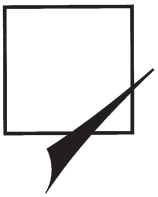
Evaluation Systems: What Are They and Why Study Them?

Frans L. Leeuw and Jan-Eric Furubo

Evaluation 2008 14: 157
DOI: 10.1177/1356389007087537

Evaluation

DOI: 10.1177/1356389007087537
Vol 14(2): 157–169



Evaluation Systems

What Are They and Why Study Them?

FRANS L. LEEUW
University of Maastricht, and Research and Documentation Center, Ministry of Justice, The Netherlands

JAN-ERIC FURUBO
Ministry of Finance, Sweden

Increasingly, in the world of evaluation, 'systems of evaluation' have been developed. This article outlines four criteria that help characterize such systems. One criterion is the existence of a distinctive epistemological perspective; another is that, in order to be labelled a system, evaluation activities are carried out by evaluators within organizational structures and institutions and not only (or largely) by 'lonely' or sole-trader evaluators. Permanence is the third criterion and the fourth is that there is a focus on the intended use of results of evaluations. Examples of systems are the performance-monitoring system, the 'experimentalist' system and the evaluation-accreditation system. Several problematic aspects of these systems are described, making it relevant to study them. One of these is the danger that evaluation systems breed (new) evaluation systems. Another problem is that these systems may produce largely routinized information relevant for day-to-day practices and single-loop learning processes, but which is of little relevance for fundamental reassessments of policies and programmes.

KEYWORDS: accreditation and evaluation; evaluation approach; experiments; inspection; institutionalization of evaluation; system of evaluation

Background

Over the last 15 to 20 years, governments and other (public sector) organizations have been paying much more attention to evaluation activities. According to some, evaluation is a 'growth industry' (Leeuw, 2001). Based on information from the *International Atlas of Evaluation* (Furubo et al., 2002), it is clear that not only has the number of evaluations carried out on behalf of governments and other public sector organizations been rapidly increasing, but also professional (evaluation) societies are now to be found in almost every European country, the USA and Canada

and in a number of African and Australasian countries. The *Atlas* also shows that (national) audit offices and parliaments have become more involved in 'evaluative practices'. Not only countries are active players on the evaluation market, but also supranational organizations like the EU and the World Bank and the like.

Rist and Stame (2006) argue that it is no longer wise to refer to 'single [evaluation] studies' that are produced, but to 'streams of studies and data': systematic reviews, syntheses, information systems, systems of good practices, m(onitoring) and e(valuation) systems, performance monitoring, inspection and oversight, repositories of evaluation results and observatories.[1]

In his keynote address during the 2006 UKES/EES Evaluation Conference in London, Elliot Stern made an important remark about the actual development of the role of evaluation within society. One of the pertinent issues he raised was the extent to which evaluators will be able to continue to speak 'truth to power' when evaluation is becoming such an important player in the public sector (Wildavsky, 1979). Stern's point is important because it is known from the field of policy analysis that success can breed its own failure. For example, Radin (2000) showed that, during the 1970s, policy analysts in the USA appeared to be overconcerned with their own needs and 'shopped for clients'.

In this article, we respond to Stern's request to reflect upon developments within evaluation and we do that from *one perspective*, i.e. that of the role of 'systems of evaluation'. Sociologists teach us that when fields of study and research develop (rapidly), 'social systems' often emerge that have their own 'logic'. This triggers the following questions:

- When can a set of evaluative activities be called an 'evaluation system'?
- Which systems can be detected?
- What are important characteristics of these systems?
- Why is it important to reflect upon them?

These are the questions we will address.

When can a Set of Evaluative Activities be Called an Evaluation System?: Four Criteria

Definitions of the concept of an 'evaluation system' are not easy to find. Often this concept is used in a technical way such that an evaluation system is something like an 'educational testing system'.[2] Recently, however, Williams and Imam (2007) edited a monograph on evaluation systems in which system thinking and its relevance for the practice of evaluation are discussed. Attenborough (2007) discussed different types of (evaluative) systems like soft systems and cybernetic systems (see also Fitch, 2007). According to Williams and Imam, thinking in terms of (evaluation) systems helps us to understand what the boundaries are of what 'lies inside and what lies outside a particular inquiry' (i.e. evaluation) (2007: 6); it also helps us to understand that systems can only exist in reference to other systems and their boundaries (e.g. the client system, the practitioner's system, the evaluators' system) (Hummelbrunner, 2007).

Of a more down to earth nature is the description that Furubo and Sandahl (2002) have presented. They see as an indicator of a mature evaluation culture 'permanent arrangements or systems whereby evaluation initiatives are commissioned to different evaluators and at the same time the evaluations conducted are put to suitable use'. To put it slightly differently: when evaluations are no longer commissioned and conducted on an ad hoc basis but through more permanent arrangements, which aim to guarantee, in advance, the supply of evaluative information, it is then wise to think in terms of evaluation systems.

We would therefore apply the following four criteria in labelling a set of evaluative activities as a system.

Criterion 1: A Distinctive Epistemological Perspective

When we refer to something as an evaluation system, the activities carried out have to be characterized in terms of a certain *cultural-cognitive perspective* (Scott, 2001). This means that there should be some agreement among the players involved about what they are actually doing and why they are doing it. For evaluation activities to be categorized as a system they need to be carried out in a *recognizable way*, based on a shared *epistemology* of what makes something an evaluative activity. Bawden (2007: 38) refers to the importance of an underlying epistemology, if one wants to detect evaluations systems. The *type of knowledge* produced is part of this (Nutley et al., 2003). Is it knowledge about what works in society, is it knowledge about why it works (or doesn't), is it knowledge of a more procedural nature, etc.? The importance of this element can be highlighted by juxtaposing two examples: performance auditing and experimental and quasi-experimental evaluations. The type of knowledge performance auditors produce (e.g. certification, compliance to rules, regulations and procedures) differs from what experimentalists produce (e.g. effectiveness, causal analysis), while both have textbooks, guidelines, websites and champions that articulate the underlying epistemology.[3]

Criterion 2: Organizational Responsibility

The second criterion is that the evaluation activities are carried out by organizations and institutions and not only (or largely) by 'lonely' or sole-trading evaluators. More precisely, it is important that organizational entities like national governments or (not-for profit) institutions have established or sponsored specific entities to undertake (or commission) the evaluations, audits or inspections. However, in order to make it a system there must be more than *one* active organization. Alongside the producer of evaluative knowledge, there must be at least one *other organization* that requests this information and that strives to use the findings.

Criterion 3: Permanence

To be labelled as an evaluation system, there should be a certain *permanence or history in the activities involved*: they are part of something ongoing. This also means that there will be a tendency to replace ad hoc initiatives and ad hoc

organizations with activities and organizations planned in advance that have a more permanent character.

Permanence also implies that there should be a certain volume of activities taking place over time. When evaluative activities only lead to a few reports produced by a few persons over a few years, there is a low score on the criterion of permanence. Permanence also includes things like specific publication outlets, relationships with mass media and links with appropriate professional societies.

Criterion 4: A Focus on the Intended Use of Evaluations

The fourth criterion is that the information from evaluative activities is (institutionally) linked to decision and implementation processes. This can be the planning process of a government department, universities or the World Bank, but it can also be the governmental budget process or adjustments of the curriculum of schools and universities. Therefore *evaluative activities* are planned in advance, taking into account the point at which the information should be submitted to decision-makers – preferably in an institutionalized way. In some cases it may be demanded that different kinds of information are to be delivered at certain intervals, for example annually or quarterly, to be useful during a specific phase of decision-making. In other cases, like accreditation, the results from evaluations would be employed as the single most important determinant of the decision whether to accredit. Here one might speak about targeted use of evaluations.

These four criteria help us to describe different systems of evaluations. It should be noted that when we describe evaluation systems existing in reality, there will be differences in terms of how 'hard', i.e. institutionalized, these systems are. The hardest, i.e. the most institutionalized, are able to provide information on an ongoing basis about the progress of (predetermined, knowledge-driven) parameters based on predetermined measurement instruments during fixed periods, while the information is intended to be used in a targeted way. 'Softer' systems do not have all this, but should have at least permanence in terms of resources and a distinct knowledge perspective from which they work.

Before describing the evaluation systems we believe are currently 'in place' in western societies, it is important to say something about the difference between evaluation paradigms – sometimes also called 'evaluation traditions' – and evaluation systems (Mertens, 2005). Although evaluations systems can be based on a particular evaluation paradigm, theory or 'tradition', the most important difference is that systems are characterized in terms of organizational capacity, sustainability, money, power and interactions with clients, stakeholders and user systems. As Alkin (2004) shows, in a number of cases evaluation traditions or theories are based on or driven by one or only a few *individual persons*, whose work has not been institutionalized. In these circumstances one cannot speak of an 'evaluation system'. However, sometimes an evaluation theory or tradition almost equals an evaluation system, as is the case with regard to the Campbell Collaboration, with its focus on experiments and quasi-experiments as drivers of evaluative knowledge.

Evaluation Systems in the Western World

Based on the *International Atlas of Evaluation* (Furubo et al., 2002) and on more recent information from symposia,[4] studies on the roots of evaluation (Alkin, 2004), workshops and other (informal) documents and meetings,[5] we present the following preliminary *overview of evaluation systems* to be found in (mainly western) countries and in organizations such as the World Bank. However, our inventory also is an invitation to readers to criticize, compare, contrast and/or add to our perspective. The overview comprises:

- the system of performance monitoring;
- the systems of performance audit, inspection and oversight;
- the system of (quasi-)experimental evaluations and the evidence-based policy movement;
- the accreditation and evaluation system;
- the monitoring and evaluation system.

The System of Performance Monitoring

An important impetus to the development of this evaluation system has been the many different ideas and reform projects within political and administrative systems under the heading of New Public Management. The Government Performance Review Act (GPRA) in the USA, the VBTB in Holland (Van beleidsbegroting tot beleidsverantwoording), the ordinance on annual reports (förordningen om årsredovisning och budgetunderlag) in Sweden and many other similar activities in other OECD countries are now well known. These activities are often sponsored by ministries of finance and with audit offices as checking agents. There also is a global 'drift' towards these activities due to the pressure of organizations like OECD. Performance monitoring systems have been implemented at different levels of government (e.g. within ministries, cities, regions and semi-independent organizations like QUANGOs, quasi-autonomous non-governmental organizations). Results-based management is a concept that is linked to performance monitoring; performance monitoring enables organizations to check to what extent the results agreed upon have been realized.

Performance monitoring can nowadays be found in all industrialized countries. An important underlying assumption is that government and society will benefit if performance measurement systems are put in place and 'do their job'. Another assumption is that, once the performance data are available, politicians, parliamentarians, managers and leaders do 'work' with them, use the information and built decisions on them. A third assumption is that there are no (important) unintended side effects, like the 'performance paradox' (Van Thiel and Leeuw, 2002).

The System of (Performance) Audit, Inspection and Oversight

A strongly connected development has been the expansion of auditing. At the level of organizations itself, but also at the level of national audit offices, the

performance audit or value for money (VFM) audit nowadays is prevalent. Supranational organizations also tend to embrace this type of activity. Power (1997) shows how widely this system has been implemented within organizations, both in the public and the private sector. Largely the focus is on auditing management and management information, but sometimes performance auditors dig deeper and go for more substantive investigations.

The growing role of audit is also an important part of the ideas labelled New Public Management (NPM) (Barzelay, 1996; Leeuw, 1996; Power, 1997). The common denominator for inspection, audit and oversight is dualistic: there is someone or some organization producing or delivering something and there is someone else or some other organization checking and controlling this process and its outcomes (usually on behalf of the taxpayer). The importance of such checking has been increased due to the NPM movement. The notion of increased freedom for managers has to be accompanied by increased scrutiny of how managers handle these 'new' levels of freedom.

This kind of scrutiny shares a salient feature which separates it from other evaluative systems. Audit, inspection and oversight have an institutional – or even an individual – perspective. The fundamental question is how *somebody* (e.g. the head of an organization) has carried out his or her task. It is not the *intervention or the policy programme as such* that is the focus of this type of evaluative work, but it is the way in which *this somebody or this organization* has acted. Pollitt points out that studies within a Supreme Audit institution '[are] bound to be done from a perspective of guardianship and control, with a basic role of holding public bodies to account for the expenditure of public funds' (Pollitt et al., 1999; see also Flint, 1988; Power, 1997). The same is largely true for inspectorates and other 'oversight' bodies. However, audit offices have a broader mandate than inspection and oversight bodies. They are often limited to a specific policy sector (e.g. education, health, work conditions).

A recent Belgian study (Put, 2005) showed that the cultural-cognitive norms applied by performance auditors working within the National Audit Offices in the UK and the Netherlands are basically the same. That is also true for the type of knowledge they produce. Most audit work looks into conditions that are considered to be relevant for making policies and administrations more effective and efficient. However, explicit empirical knowledge of effectiveness and efficiency is hardly ever demonstrated in reports by performance auditors in these countries (Pollitt et al., 1999; Put, 2005; Schwartz, 1999). Mayne takes the discussion a step further and in a more normative direction when he argues that performance audit should avoid that kind of issue, i.e. that performance audits should not assess impact and effectiveness (Mayne, 2006).

The System of (Quasi-)Experimental Evaluations and the Evidence-Based Policy Movement

The experimentalists in evaluation have a long history, going back to the 1930s in the field of experimental sociology and psychology in the USA and to the 1950s in the UK in the field of crime reduction (Leeuw, 2005). Later, in the

mid-1960s in the USA, during the 'Great Society' programmes, this type of evaluation started to blossom. Longitudinal experiments such as one on income maintenance programmes (Oakley, 2000, 2005) were carried out. Over the last 15 years there has again been a boom, now affiliated with the 'evidence-based policy' movement, the 'what works' tradition and the Campbell Collaboration. This organization considers the experimental design as the single most important evaluation design to use. The Collaboration is a worldwide network of researchers and commissioners and sponsors of research actively involved in producing evidence along the lines of experiments and quasi-experiments. It is linked to the Cochrane society, which is older and focused on medical evaluations. The Cochrane criteria of reviews and randomized control trials are widely diffused; 250,000 experiments have been carried out, while some 11,000 have been conducted in the social, economic and behavioural sciences. In the Campbell and Cochrane repositories, information about experiments and their systematized reviews can be found.

The underlying epistemology can be found in numerous publications. The *Maryland Scientific Methods scale* – which is often used within, for example, criminological research – illuminates its epistemology. It distinguishes between five types of evaluation designs, with correlation at the bottom and the experimental method (randomized control trial) at the top.

The Accreditation and Evaluation System

The symbiosis of accreditation and evaluation has become an important part of the evaluative landscape in a number of policy fields and organizations. Examples are education, prison management, quality control, energy conservation, hospitals, broadcasting organizations, NGOs and many more. Today accreditation is very well known in Europe and the USA. Accreditation companies, sometimes public, sometimes private, undertake the work. Peer review, 'visitation' activities and document analysis are parts of the epistemology underlying the work done by evaluator-accreditors. Interestingly, accreditation often starts with self-evaluation by the organizations that strive for accreditation. The self-evaluation is then reviewed by the evaluator-accreditor; if the quality of the self-evaluation is considered satisfactory, often only a marginal new and empirical assessment of the organization's work is sufficient to grant the organization its accreditation. Increasingly attention is paid to risk-assessments and risk-management. The evaluator-accreditor thereby analyses what are the major risks for schools or hospitals and then decides on which to focus his attention. The lower the likelihood of (perceived) risks, the less often on-site investigations will take place. Self-evaluations form a part of this work. Regularly a review of these self-evaluations is carried out. The more critical this review, the larger the probability that the evaluator (or 'inspector') will start his own data collection. Sometimes a 'single information – single audit' approach is also followed.

The Monitoring and Evaluation System

M&E systems are to be found in many different fields, like development aid, health, crime and justice and social welfare. Kusek and Rist (2004) have

articulated the underlying philosophy:

> Building an M&E system essentially adds the fourth leg to the governance chair. Typically and traditionally, governments have built budget systems, human resource systems, and auditing systems as the three legs of a governance system. But what has been missing has been the feedback system on the outcomes and consequences of government actions. This is what building an M&E system brings as an additional public sector management tool.

Organizations such as the World Bank, ministries for development aid and NGOs are behind these initiatives in the world of development aid policies. There are also specialized institutions to carry out the work. In the Netherlands, for example, they are active in the field of social policies, crime and delinquency policies, the labour market and the field of integration of minorities within society. For the topic of integration of minorities in Dutch society alone, some 19 monitoring instruments are currently implemented that collect and feed back information to policy-makers (Kulu-Glasgow et al., 2007). In Sweden there are similar special 'research bodies' doing this work. Also, slowly but steadily, national audit offices are active, in the sense that they monitor and audit the M&E systems.

It should be noted that monitoring and evaluation on its own does not warrant reference to a 'system'. Two things are necessary to make it a system. First, in a number of countries and organizations a development has taken place in which special bodies, often at an arm's length from implementation and performance monitoring, have been created that have as a specific task to undertake the M&E work. Both in Sweden and the Netherlands such bodies can be located. In both countries examples can be found in areas such as crime prevention, social welfare, international assistance, labour-market and economic growth policies. These institutions have often a more autonomous position than other governmental bodies and have in some cases close links to university institutions. They have evaluation as their main task, but also undertake other forms of research.

These bodies are semi-administrative and semi-academic in nature, which gives them a special position within the bureaucracy. Second, these bodies in particular are focused on monitoring outcomes, impacts and effects of policies and programmes. In earlier years, monitoring activities were largely focused on questions about efficiency or 'economy', while outcomes and impacts were less often 'monitored'.

With regard to the underlying epistemology, we believe it is safe to argue that it resembles a social engineering approach. It is based on at least two assumptions. The first is that it is possible to acquire knowledge about earlier interventions in a way that is in a rather direct sense useful for decisions about the scale and scope of future interventions. The second is that capacity building at a macro (country) level is believed to increase the learning capacity of the (public) sector in these countries.

Why Study Evaluation Systems?

Evaluation is a characteristic of modern states and, although it took some time to develop, it is now almost considered a 'natural' phenomenon. Common sense leads us to believe that evaluation produces relevant knowledge, which helps us to

make better decisions. However, fundamental questions can and should be raised about evaluation. This article has so far focused on one aspect of the present situation: *the development of 'systems of evaluation'*. At least four hypotheses have to be addressed in a more empirical discussion about evaluation systems and their role within society, as follows.

1. Evaluation systems produce largely routinized information relevant for day-to-day practices and single-loop learning but of little relevance for fundamental reassessments and double-loop learning Most of the evaluation systems are embedded in administrative structures. The marriage between administration and evaluation has meant that administrators and evaluators have joined forces to ensure that politics can be characterized as a type of *administrative scholarliness*. This has resulted in combinations of evaluators and administrators becoming the interpreters of the politicians' need for information. This, in turn, has led to an expansion of information believed to be useful in administrative decision processes. In Europe we have much more information about costs, outputs, internal quality and process parameters underlying the development and implementation of policies than 20 or 30 years ago. But simultaneously, the well-known phenomenon of the 'performance paradox' is around, indicating that more of this type of information does not guarantee (more) effective policies (Van Thiel and Leeuw, 2002). Some are even more critical and say that we have perhaps even less knowledge available which helps politicians or others question the assumptions behind fundamental policies.

2. Evaluation systems are largely used as providers of (procedural) assurance Evaluation systems produce knowledge about the ways in which policies are implemented and about their accomplishments. A society that has implemented systems of evaluation creates a feeling of assurance amongst its members and its leaders: 'the *inspectorate* is looking into the problem'; 'the *evaluation institute* X has found that…'; 'this new university has been *accredited and will act accordingly*'. This providing of assurance information is almost becoming an economic good that deserves a systematic, industry-type production.

Evaluation systems create not only a sense of assurance and of having things under control, but also an effective management of reputational risks for top-level civil servants, heads of QUANGOs and NGOs, and sometimes even politicians. When systems of evaluation have been implemented, top-level management can no longer be accused of negligence, laid-back or irresponsible behaviour, because management can persuasively argue that 'it has done everything to prevent this problem' or that 'it has done everything to monitor these processes'. The more fundamental question whether or not 'policy mistakes' and 'policy disasters' have been prevented or, when they have happened, have resulted in minimal costs, is not always on the agenda.

3. Evaluation systems produce information that confirms rather than questions policies From research into public administration, the phenomenon of 'goal displacement' is well known. It has different dimensions. One dimension is that:

> … public agencies are frequently evaluated based on the *outputs* they produce. Agency outputs (e.g., criminal cases solved, inspections done) are easier to measure than the

> actual contributions agencies make to desired social outcomes (e.g., preventing workplace discrimination, protecting the environment). When agency performance is evaluated in terms of numerical outputs, bureaucrats have an incentive to maximize outputs, regardless of whether maximizing outputs is the preferred strategy for achieving desired social outcomes. (Bothe and Meier, 2000)

Bothe and Meier label this a form of goal displacement.

Another dimension of goal displacement is that evaluations risk producing information that is based on the assumptions about how the policy or programme works; they then consider corroboration of these assumptions as a positive indicator of the effectiveness of these programmes. The following example illustrates this.

If policy-makers use information diffusion as a policy instrument to convince the public (e.g. house-owners) about the need for more energy-conservation measures, through producing and dissemination information that energy conservation gives society the best of both worlds (less costs and lower environmental pollution), the information campaign to be evaluated is also based on this assumption. Next, evaluators measure the impact of this campaign in terms of the number of people that *know more* about energy conservation. *If* evaluators then use *these* changes in knowledge as an *indicator* of how successful the campaign has been, this can also be considered 'goal displacement'. Instead of finding out to what extent house-owners indeed are engaged in 'energy conservation *behaviour*', the evaluators believe that information about a proxy indicator is sufficient.

Evaluation systems may tend to produce more evaluations of this kind, therefore reducing the possibility of questioning the assumptions on which the programmes are based and which underlie the impact of the programmes. To some extent this phenomenon is linked to Friedman's concept of *designed blindness*. This refers to the problem that a strong belief in the assumed validity of a programme theory makes the actors concerned unintentionally blind to the side-effects or negative effects of their actions based on this programme theory (Friedman, 2001).

4. Evaluation systems breed evaluation (systems) Budget maximization was a phenomenon detected by public choice economists like Downs (1957) and Niskanen (1994). Put simply, it is the idea that, where entrepreneurs aim to realize financial profits, public sector officials (not being able to make profits) work towards bigger budgets, bigger limousines, bigger rooms, more 'turf' and better secretaries (see Mueller, 2003). (There is, of course, also a body of research that suggests that the managers of larger private sector corporations also tend to maximize status, turnover and prestige – hence the motive for mergers and acquisitions – rather than profits.) Radin (2000) discovered that, within the world of policy analysis, a similar 'mechanism' worked. It dealt with the pressure to make sure that policy analyst organizations continue to exist, keep their personnel busy, keep their reputations alive and – therefore – start to get involved in shopping-for-clients behaviour. Critical analysis that once characterized independent thinking came under pressure, the more policy analysts became dependent upon (governmental) funds. Given the boom in evaluation (systems), a similar thing might be happening to evaluation systems.

Conclusions

We draw a number of conclusions. First, the evaluation community should invest (more) in reflecting upon what now appears to be adequate for the (further) improvement of the world of evaluation, but what might – in the not too distant future – begin to look ugly. Creating and sustaining systems of evaluation with their own incentives that can lead to perverse effects, such as 'shopping for clients', 'tunnel vision' and predictability of what is being studied, will make it less possible to 'speak truth to power'. Success may breed its own failure. We therefore urge the evaluation community to get involved in debates on and studies about 'systems of evaluation'.

Second, how evaluation systems are able to 'make' and 'create' certain blindnesses needs to be explored more. Research on 'dramaturgical compliance' in the world of (peer) reviewed committees in academia, hospitals and municipalities (Goodlad, 2000; Power, 1997) helps us to understand how systems influence frames of reference and behaviour.

Finally, we hope that this article will not only lead to a discussion of the pros and cons of evaluation systems but will also encourage other researchers to consider, in more depth than we did, which evaluation systems currently exist and how they behave.

Notes

This contribution was based on the keynote speech by Leeuw during the 2nd Conference of the Swedish Evaluation Society in Stockholm in October 2005 and the introductory speech by Furubo during the same conference (see www.svuf.nu). We thank an anonymous reviewer who suggested a number of improvements, such as discussing differences between evaluation paradigms and evaluation systems, and who also presented us with references to systems thinking.

1. They do not produce any data themselves but their mission is to bring together, prepare and disseminate existing statistical data to monitor specific policy problems in order to improve knowledge about the issue.
2. See examples like www.nesinc.com (National Evaluation Systems, Inc.), an Evaluation System for Courses and Instruction or a 'total quality management' system. These 'systems' are implemented within organizations to help undertake evaluative activities. We will use the word 'evaluation system' on a more institutional level.
3. Linked to this criterion is the point that sometimes the evaluative information can be more or less strongly associated with precise demands regarding the content of the information to be produced. Sometimes evaluators are asked in advance to provide certain data, like cost–benefit studies or ex ante impact assessments at certain times.
4. For example, the Swedish Evaluation Society devoted its 2005 conference to this topic.
5. Some examples are the INTEVAL Working group led by Ray C. Rist, a workshop organized in Sweden by Ove Karlsson to develop a European equivalent of Marvin C. Alkin's book on *Roots of Evaluation* (2004), the UKES–EES Conference (London, 2006), an international meeting organized by the Dutch Evaluation Society and the Netherlands National Audit Office (2007).

References

Alkin, M. C., ed. (2004) *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: SAGE.

Attenborough, K. (2007) 'Soft Systems in a Hardening World: Evaluating Urban Regeneration', in B. Williams and I. Imam (eds) *Systems Concepts in Evaluation: An Expert Anthology*, pp. 75–89. Point Reyes: EdgePress of Inverness.

Barzelay, M. (1996) 'Performance Audit and the New Public Management: Changing Roles and Strategies of Central Audit Institutions', in *Performance Auditing and the Modernisation of Government*, pp. 15–57. Paris: OECD.

Bawden, R. (2007) 'A Systemic Evaluation of an Agricultural Development: A Focus on the Worldview Challenge', in B. Williams and I. Imam (eds) *Systems Concepts in Evaluation: An Expert Anthology*, pp. 35–47. Point Reyes: EdgePress of Inverness.

Bothe, J. and K. J. Meier (2000) 'Assessing the Motivation for Organizational Cheating', *Public Administration Review* 60 (March/April): 173–82.

Downs, A. (1957) *Inside Bureaucracy*. San Francisco, CA: Little, Brown & Co. and Rand Corporation.

Fitch, D. (2007) 'A Cybernetic Evaluation of Organizational Information Systems', in B. Williams and I. Imam (eds) *Systems Concepts in Evaluation: An Expert Anthology*, pp. 65–75. Point Reyes: EdgePress of Inverness.

Flint, D. (1988) *Philosophy and Principles of Auditing. An Introduction*. Houndmills: Macmillan Education Ltd.

Friedman, V. J. (2001) 'Designed Blindness: An Action Science Perspective on Program Theory Evaluation', *American Journal of Evaluation* 22(2): 161–81.

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

Furubo, J.-E., R. C. Rist and R. Sandahl (eds) *International Atlas of Evaluation. London*: Transaction Publishers.

Goodlad, S. (2000) 'Benchmarks and Templates: Some Notes and Queries Form a Sceptic', in H. Smith et al. (eds) *Benchmarking and Threshold Standards in Higher Education*. London: Kogan Page.

Hummelbrunner, R. (2007) 'Systemic Evaluation in the Field of Regional Development', in B. Williams and I. Imam (eds) *Systems Concepts in Evaluation: An Expert Anthology*, pp. 161–81. Point Reyes: EdgePress of Inverness.

Imam, I., A. LaGoy and B. Williams (2007) 'Introduction', in B. Williams and I. Imam (eds) *Systems Concepts in Evaluation: An Expert Anthology*. Point Reyes: EdgePress of Inverness.

Kulu-Glasgow, I., F. L. Leeuw, E. Uiters and R. V. Bijl (2007) *Integratiebeleid rijksoverheid onderzocht: Een synthese van resultaten uit evaluatie- en monitoring-onderzoek, 2003–2006*. Amsterdam: Boom Legal Publishers.

Kusek, J. Z. and R. C. Rist (2004) *Ten Steps to a Results-Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. Washington, DC: World Bank.

Leeuw, F. L. (1996) 'Performance Auditing, New Public Management and Performance Improvement: Questions and Challenges', in *Performance Auditing and the Modernisation of Government*, pp. 105–16. Paris: OECD.

Leeuw, F. L. (2001) 'Evaluation in Europe: Challenges to a Growth Industry', *Evaluation* 8(1): 5–13.

Leeuw, F. L. (2005) 'Trends and Developments in Program Evaluation in General and Criminal Justice Programs in Particular', *European Journal on Criminal Policy and Research* 11: 18–35.

Mayne, J. (2006) 'Audit and Evaluation in Public Management: Challenges, Reforms and Different Roles', *Canadian Journal of Program Evaluation* 21(1): 11–45.
Mertens, D. M. (2005) *Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods*, 2nd edn. Thousand Oaks, CA: SAGE.
Mueller, D. C. (2003) *Public Choice III*. Cambridge: Cambridge University Press.
Niskanen, W. N. (1994) *Bureaucracy and Public Economics*. Fairfax, VA: Locke Institute.
Nutley, S., I. Walter and H. Davies (2003) 'From Knowing to Doing: A Framework for Understanding the Evidence-into-Practice Agenda', *Evaluation* 9(2): 125–49.
Oakley, A. (2000) *Experiments in Knowing: Gender and Method in the Social Sciences*. Cambridge: Polity Press.
Oakley, A. (2005) 'The Politics of Evidence and Methodology: Lessons from the EPPI-Centre', *Evidence and Policy* 1(1): 5–32.
Pollitt, C., X. Girre, J. Lonsdale, R. Mul, H. Summa and M. Waerness (1999) *Performance Audit and Public Management in Five Countries*. Oxford and New York: Oxford University Press.
Power, M. (1997) *The Audit Society: Rituals of Verification*. Oxford and New York: Oxford University Press.
Put, V. (2005) 'Normen in performance audits van rekenkamers: een casestudie bij de Algemene Rekenkamer en het National Audit Office', unpublished doctoral dissertation. Leuven, Belgium: University of Leuven.
Radin, B. A. (2000) *Beyond Machiavelli: Policy Analysis Comes of Age*. Washington, DC: Georgetown University Press.
Rist, R. and N. Stame, eds (2006) *From Studies to Streams: Managing Evaluative Systems*. London: Transaction Publishers.
Schwartz, R. (1999) 'Coping with the Effectiveness Dilemma: Strategies Adopted by State Auditors', *International Review of Administrative Sciences* 65(4): 511–26.
Scott, W. R. (2001) *Institutions and Organizations*. Thousand Oaks, CA: SAGE.
Summa, H. (1999) 'Definitions and Frameworks', in C. Pollitt et al. *Performance Audit and Public Management in Five Countries*, pp. 11–29. Oxford and New York: Oxford University Press.
Van Thiel, S. and F. L. Leeuw (2002) 'The Performance Paradox in the Public Sector', *Public Productivity and Management Review* 25(3): 267–81.
Wildavsky, A. (1979) *Speaking Truth to Power: The Art and Craft of Policy Analysis*. Boston, MA: Little, Brown.
Williams, B. and I. Imam, eds (2007) *Systems Concepts in Evaluation: An Expert Anthology*. Point Reyes: EdgePress of Inverness.

FRANS LEEUW is professor of law, public policy and social science research at Maastricht University, The Netherlands and Director of the Research and Documentation Center, Ministry of Justice, WODC, Den Haag, The Netherlands. [email: frans.leeuw@metajur.unimaas.nl]

JAN-ERIC FURUBO is at present the main-secretary in a governmental committee within the Ministry of Finance considering public administration and the role of evaluation versus other forms of knowledge in political decision making. [email: furukem@bostream.nu]