REVISED DRAFT

New Public Management and Evaluation

Frans-Bauke van der Meer

Introduction

In the last four decades or so, governments and governmental bodies in 'western' countries have come under increasing pressure, both societal and financial in nature. The quest for efficiency, efficacy, responsiveness, flexibility and accountability of government is a dominant one. As outlined in the introductory chapter, this has given rise to numerous innovations both in public organization and in policymaking and implementation. These innovations aimed at improving government in some respects in order to accommodate 'external' and 'internal' pressures.

Not all these innovations are considered successful. Some are replaced by others within a few years time, in which case a change of direction (e.g. on a centralization – decentralization dimension) is not uncommon (Pollit and Bouckaert 2000: 149-152). In other cases innovations are supplemented with other ones, e.g. autonomization followed by a growing intensity of central control (Bekkers 1998; Van Thiel and Pollitt, this volume, chapter 4). On the other hand there are successes, at least in the eyes of some actors involved (see Pollitt in chapter 2 and Kickert in chapter 3, this volume). Such 'best practices', however, may be copied elsewhere with mixed results (e.g. Hakvoort and Klaassen, this volume, chapter 7). Moreover authorities and agencies are frequently criticized (e.g. in cases of a disaster) for procedures and practices they have taken for granted for long. It may be concluded that – although many procedures and innovations are considered 'necessary' and 'unavoidable' – not enough is known about their working and impact (Cf. Teisman and Van Buren in chapter 11 of this volume).

In this chapter I will discuss what evaluation does and can contribute to insight in the dynamics and effects of (new) public management reforms. To do so, I pose three key questions:

1. How is evaluation actually applied in relation to NPM-reforms?
2. How are evaluation results used and how can their impact be understood?
3. How can evaluations in and of NPM-reforms be improved?

The theoretical framework that I will present to help answering the second question will also be used to explain the state of affairs in relation to the first question. Moreover, it will be used, in combination with empirical observations to support the conclusions with respect to question 3.

Before going into the three questions in turn I present some observations on the broader context of reforms in the public domain. In my view this is relevant for the present subject, because evaluation of specific measures or of specific (classes of) reforms should take into account relevant boundary conditions and goals or requirements that are not explicitly included in the focal measures.

Challenges for public management

NPM-type reforms in the public domain are quite widespread (Kickert 1997; Pollitt and Bouckaert 2000). Although the label NPM refers to a rather heterogeneous set of innovations and practices (see chapter 1), it can be said to comprise the utilization of private sector management techniques and organizational forms in the public sector. NPM innovations seek to improve efficiency and quality of public administration and service delivery by decentral autonomy and output control, frequently regulated by management contracts. The underlying rationale seems to be that an incentive structure that rewards efficient production and measurable output enhances motivation and creativity of individuals and groups and can replace much efficiency reducing hierarchical control. Moreover the decentralized nature of the new arrangements is thought to enhance flexibility, responsiveness and tailor-made solutions: a demand orientation, which is presupposed to be the main guide of private sector management. Taken together NPM reforms are expected to trigger a better use of public sector resources.

NPM reforms are by no means the only type of reforms with which public and semi-public agencies are occupied. Nor is the underlying rationale of NPM the only rationale about improvement in the public domain. A second type of innovation focuses not so much on government or public sector production per se but on joint production of policies, public works and services with private sector enterprises and/or societal organizations (Teisman and Verheij 1996; Tops and Weterings 1998). The aims are in a sense similar to those of NPM innovations, viz. improving efficiency and quality of public production. But here (semi-)public agencies are not the sole producers. Here too, there is a parallel with private sector strategies, where organizations initiate joint enterprises to realize projects, develop new technology or penetrate new markets. The main rationale, however, is quite different from the NPM type. The problem is not that the public sector uses its resources in a sub-optimal way, but that government lacks key knowledge, financial and other resources to realize many projects and

solutions to societal problems. Government is only one among many actors in a societal network of interdependencies. Only by cooperation effective results can be hoped for.

A third set of new strategies in public management seeks to involve citizens and societal organizations having an interest in certain policy issues or domains. This participation (Edelenbos 2000) may include the phase of problem definition, generation of possible solutions, a role of citizens and societal organizations in policy implementation and evaluation, and a shift in responsibility towards the citizens or private initiative. By definition there is some overlap with the second type op public management innovations since here too, non-governmental actors are involved in the production of policy and policy outcomes. However, there are other rationales for such participative strategies. One of them starts from the notion that societal problems often cannot be identified in an objective and unambiguous way. The public task, then, is not so much to solve 'the' problem, but to search for ways to accommodate different problem definitions at the same time (Teisman and Verheij 1996). By involving individuals and groups having an interest in at least some of these definitions, the searching process is part of its solution. In and by interaction with other actors (among which governmental ones) actors may change or broaden their problem definitions and/or come to conceive of new solutions. Thus, eventual outcomes may in part be created by and gain the support of stakeholders. The role of government is mainly a facilitating one in producing solutions and support (De Bruijn et al. 1998).

Now, it is important to notice that these different rationales and steering practices, as well as classical hierarchical approaches are applied simultaneously, often by the same actors, sometimes within a single policy initiative.[1] Thus, is not only worthwhile to ask how successful individual public management innovations are and how valid their rationales are but it is also interesting to see how different types of reforms influence each other and how their rationales interact in practice, for these interaction influence behaviors and outcomes. Therefore, evaluation should take account of them.

Evaluation as instrument in NPM practices

In the framework of NPM-like innovations evaluations play an important role. Assessment of performance or output and comparing it to initial targets or contractual commitments is cen-

[1] For example, in the new Dutch law on care ('Wet Maatschappelijke Ondersteuning') all four rationales can be found:

- hierarchy: the municipalities are made responsible by law for societal care
- co-production: care should be delivered as much as possible by private organizations
- participation: care clients should participate in formulating care policies; clients should take responsibility for organizing their own care as much as possible
- new public management: care delivering organizations are contracted and evaluated according to agreed performance measures

tral to the philosophy of this type of approach. Moreover, the felt need for public accountability of (autonomized) agencies has given rise to numerous review procedures and control mechanisms, which also involve the collection and judgment of data on performance and results (see e.g. Sanderson 2001, who describes local practices in the UK). Such practices contribute to an 'evaluative state' (Sanderson 2001: 303) or 'audit explosion' (Power, 1994) in which performance measurement and accountability constitute the dominant perspective (see e.g. this volume, chapter 4).

Although there are differences between countries with respect to the quantity and specific characteristics of performance evaluations (e.g. Christensen et al 2003), there is a clear tendency, at least in the Anglo-Saxon and Northern European world towards more and more quantitative performance assessments. This holds both for 'internal' management reports and external audits (Pollitt et al. 1999; Wollmann 2003).

With respect to the utilization and impact of these evaluations four observations can be made.

First, they influence and direct actual behavior. Since performance indicators generally are known before the actual performance takes place, they can and will guide this performance (Van Thiel and Leeuw, 2002). If the performance of academic work is assessed by counting the number of publication in international refereed journals, the number of such publications and the time devoted to producing them will rise. In a sense, the impact precedes the actual evaluation. But there are also ex post impacts as for example the allocation of new budgets is based on evaluation results: consecutive action is constrained or enhanced or redirected by such reallocations.

Second, this type of evaluations may give rise to unintended or perverse effects. This is a consequence of the fact that performance indicators used generally are only partial operationalizations or proxies of the goals aimed at. This phenomenon is known as the performance paradox (Meyer and Gupta 1994; Van Thiel and Leeuw, 2002).

Third, frequently, especially in the relation between 'principal' and 'agent' the amount and the specificity of data produced by agents are so large that their impact is greatly reduced by the overload or ambiguity it creates for the principal (Hazeu 2000). The steering capacity of the principal may actually be reduced by the massive introduction of performance data.

Finally, the fact that evaluations in this context focus on accountability may produce 'defensive' reactions by evaluated units in the sense of denying the validity of evaluation outcomes or seeking excuses beyond control (Sanderson, 2001; Teisman and Van der Meer 2002). Such reactions reduce the opportunities for learning since the question of possible improvements is evaded.

Evaluation of NPM reforms

In view of the impressive amount of evaluations produced in relation to NPM-like management, it is remarkable that NPM-reforms as such, including the evaluation procedures involved, are evaluated far less systematically (Pollitt 1995; Pollitt and Bouckaert 2003; Forbes and Lynn 2005). Broadbent and Laughlin (1997) observe much resistance in UK government towards such evaluations, for budgetary and other reasons. Still, in the meantime there have appeared a number of studies. The edited volume by Wollmann (2003) gives an impression, although it is not always clear whether its reviews pertain to evaluation as an instrument within NPM or to evaluation of the reforms as such. In part this can be excused for it should be noted that there is no sharp demarcation between these two categories of evaluation.

There appears to be a grey zone in between in which NPM reforms are evaluated within the logic of NPM itself. For example, the success of a reform may be measured by the extent to which performance data are produced, and not so much by a related increase in efficiency or responsiveness. In part this may be due to methodical problems such as a poor measurability, or lacking ex ante data. But even apart from that there often seems to be a remarkable inclination, especially in audit institutions, to focus on administrative control and documentation in stead of on eventual results.

A study by Boyne et al. (2003) is interesting, not only because it tries to assess intended outcomes of public management innovations in the British education, health care and housing sectors, but also because it devotes attention to unintended effects with respect to equity. Interestingly, this last variable appears to be the only one with an unequivocal, and negative, result. Broad and open evaluations like this one seem to be the exception rather than the rule.

Dynamics of evaluation

After this rough sketch of practices of NPM evaluation, I now venture into a theoretical reflection about why evaluation practices are shaped in the way they are, when, why and how their results are used and how their impacts are generated. Next, I will use this theoretical framework to contribute to an understanding of the dynamics of evaluation in and of NPM (see boxes) and to develop ideas for its improvement (see following sections).

I propose to view evaluation in terms of sensemaking processes (Weick 1979, 1995). The simple and somewhat naive argument to justify this approach is that evaluation aims at sensemaking. It is about assessing what is or has been going on, how this has come about, how the findings have to be valued and about how we might think of possible improvements. But, although the previous sentence may reflect the ambition of evaluators and/or their principals,

we should realize that it may not be shared by all actors involved in the evaluation process. Some may see it as an irrelevant time and money consuming bureaucratic practice or as a threat to their autonomy or professional discretion. Some may loyally contribute data but not hope or intend to learn something from it. However, these other actors are still engaged in sensemaking. They attach meaning to the evaluation process and its outcomes. For our present purpose this is not interesting in itself. The key argument for viewing the dynamics of evaluation in terms of sensemaking processes is the thesis that the interplay of the different sensemaking processes in the actions and interactions of actors determine the shape, the outcome and the impact of evaluations. I will elaborate this thesis in a number of steps below.

First, in the stage of initiating an evaluation a large number of choices are made by commissioners and evaluators. Amongst these are substantive choices, such as: What exactly is to be evaluated? Which criteria are to be used? Which types of data are required? There are also methodological choices: How are data to be collected? How are they to be interpreted and related to each other? Finally, there are organizational or administrative choices: Who is to perform the evaluation? Who are to be involved in what roles? Who should take cognizance of the results and act upon them?

How are such choices, which constitute the 'script' for the evaluation, made? Presumably they are the result of interactions (consultation, negotiation) between the actors involved in this stage. This interaction takes place in the context of already existing frames of meaning and patterns of practice. I call such patterns of thinking and acting *repertoires* (Van der Meer 1999; Van der Meer et al. 2000). To the extent in which there is consensus among the actors involved, the definition of the evaluation will reflect the common repertoire. Where there is dissensus, negotiation, compromise and use of 'power' will come about, often resulting in vague or ambiguous formulations. But even then, the resulting design reflects sensemaking form repertoires. Agreement is reached by means of (implicit) answers that actors give to questions like: What is acceptable? What are driving forces? What opportunities are left for influence or manipulation in a later stage? Thus, sensemaking in interaction gives shape to the evaluation.

Commissioners of NPM related evaluations, who frequently are co-initiators or sponsors of the reform, and the evaluators they hire generally will have a strong believe in the rationale behind NPM reforms. Therefore, in thinking about evaluation they will for example tend to start with the idea that a result-oriented structure will evoke maximal motivation and productivity (both quantitative and qualitative). Thus, in their view evaluation should assess results and measure these in terms of success or accordance to preset goals. The agent should then be

rewarded accordingly. Within the NPM rationale this argument with respect to evaluation seems to be self-evident. From this perspective there is little reason to doubt the rationale itself, which may explain why NPM reforms are seldom explicitly evaluated. And when they are, evaluations tend to focus on the extent to which organizations involved bring NPM procedures and provisions into practice and far less on what the eventual impacts of the reforms are.

Second, in carrying out an evaluation, again sensemaking and interaction take place. In this stage typically other actors become involved. Often commissioners will be involved less, but others, such as managers and other personnel of evaluated units, clients, citizens, partners, etc. are interviewed, are requested or required to produce data, or are even asked for their own opinions and assessments as in participative or responsive evaluation (Abma 1996; Taket and White 1997; Ryan and DiStefano 2001). Obviously, these new actors have to make sense of what is asked of them, even if they take the evaluation for granted and adopt an a priori positive attitude. They do so by utilizing their own repertoire, which may differ considerably from the repertoire of the initiators of the evaluation. Moreover, the new actors will also make sense of the evaluation as such: What are its aims? How valid will the results be? What consequences may the results have? This sensemaking will of course influence their responses. This may be called 'strategic' behavior, as long as that is not synonymous with disloyality. It is clear from this argument that the evaluation is considerably influenced by sensemaking processes in this stage. It might even be said that the evaluation often is redefined during its course.[2]

Professional employees in a unit involved in some sort of contract may make – from their professional perspective - a rather sharp distinction between the quantified output criteria mentioned in the contract and 'real' quality, which is hard to measure. If so, they may consider it fully legitimate to score on the output criteria as quickly and artificially as possible in order to safe time for the real thing or to induce an 'accurate' evaluation although the indicators are invalid from a professional point of view (Green, 1999).

Also, managers of agencies will have their own impressions and make their own assessments of the functioning of their organization and its efficacy. These images may not fully correspond to the picture that emerges from the data to be included in the evaluation, for example because the evaluation data do not reflect relevant changes in external conditions as they are experienced and interpreted by the manager. This may trigger the manager to provide additional data, to present data in other ways, or to supply extensive comments to

[2] Cf. Dawson (1996, 233-262) who makes a similar analysis of the implementation of planned change

the data. However, he may also refrain from such actions if he fears that these will be interpreted as defensive window dressing by the principal.

A third step in my theoretical argument is about the impact of evaluations (Cf. Van der Meer 1999). Here again the notion of sensemaking plays a key role. The influence of evaluations and evaluation results on actual behavior (talking, rewarding, changing, acting) is determined by how they are interpreted and valued by actors involved. Besides that, the influence of evaluation on actual administrative and organizational behavior depends on the way its outcomes become embedded in the web of other factors influencing this behavior (Weiss 1990) It is important to note that in this stage generally there are far more actors involved than those who commissioned the evaluation. Although the latter may sometimes have the power to make decisions with respect to tasks, personnel, budgets, rewards, or organizational structures and procedures, they are by no means the only actors whose sensemaking matters. I have already pointed to the possibility that evaluation evokes perverse effects or defensive reactions if that seems to be useful or self-evident from the perspective of evaluated actors.[3] Moreover, evaluated actors may try to 'improve' their actions in anticipation of or reaction to evaluation. The way in which they proceed in this connection will again depend on how they interpret both their own behavior and the evaluation and its outcomes. Finally, also the commissioners of the evaluation or those with formal power to make decisions on its consequences have to make sense of the evaluation results – explicitly or implicitly - before they can act upon it.

In NPM contexts evaluations frequently focus on performance measurement and comparing quantified results with ex ante targets so as to enable accountability. Thus, evaluations may lead to reallocation of budgets or revision of contracts. These attributes may well have quite different meanings for the actors involved.[4]

A typical principal might feel the evaluation gives him control and enables him to bear an overall responsibility without the necessity of going into implementations details. He may, furthermore, feel secure even if he does not fully understand the details of the evaluation, because he holds a strong believe in the rationales behind NPM. The general picture of success or failure suffices (cf. the fact that surveillance is often delegated to independent overseeing bodies). Sensemaking and consecutive action seem straightforward, although additional information supplied (see previous box) may cause some equivocality. But if 84,7% of

[3] In fact, the labels 'perverse' and 'defensive' are interpretations and valuations based on sensemaking processes themselves. There are few actors considering themselves perverse or defensive. This precisely exemplifies the diversity of sensemaking in these processes.

[4] Cf. Schein 1996 on typical differences in 'culture' between operators, technicians and managers, linked to their tasks, knowledge and position.

Dutch passenger trains are ‘in time’, while the contract required 86,5%, that’s a failure.

For a manager of an implementing agency the evaluation results may be far more ambiguous. He may have doubts about the validity of the results, he has to think of explanations and he needs to invent internal measures and external strategies to deal with the situation shown or induced by the evaluation. In this sensemaking and strategy development process there is far more at stake than the straightforward comparison of targets and realization. It determines to a large extent how the organization will ‘learn’ from the evaluation, but it also determines the negotiation strategy with the principal, which may eventually influence the decisions the latter will take.

Professional employees are often in a position in which they do not need to react directly to evaluation outcomes. However, they are confronted with possible management measures. Their assessment of the validity and relevance of the evaluation and of the sensibility and force of the management measures will influence their behavior. They may feel forced to ‘conform’, although perhaps in ‘perverse’ ways, or they may feel their professional status and factual autonomy is large enough to virtually negate the evaluation results and/or the management measures. Also, if their ‘results’ are successful, this may enhance their immunity from critique in later episodes.

A following theoretical idea connecting the dynamics of evaluation processes to sensemaking and interaction is that third actors play a role. Evaluation is not a sensemaking process that takes place in isolation between principal and agent or between evaluator and evaluated, but in a context full of other actors, such as competitors, clients, citizens, interest groups, parliament, press, etc. Such actors may formally or informally influence the focus or criteria used in evaluations. They may also react to evaluations (or their divers follow-ups). If these third actors (are considered to) have power or are thought of as ‘important’ by the focal actors, their reactions may contribute to an indirect influence evaluations may have on these focal actors. Moreover third actors may adapt their own behavior on the basis of their interpretation of evaluation outcomes.

Suppose, some targets are set for the personal safety policy in a certain neighborhood and an evaluation report shows that safety has increased, while at the same time many inhabitants of the neighborhood feel less secure than they did a year ago. They may protest and seek media coverage for their problems. Or they may realize that they can influence (i.e. make more realistic from their point of view) performance indicators, e.g. by reporting incidents more frequently to the police. The first reaction may for example trigger a decision to include subjective experience of inhabitants in next year’s performance review (as is actually done in the

Rotterdam safety monitor). The second reaction may lead to a lower score in the following year, which may trigger measures that are not actually warranted by worse performance. (Cf. Van Thiel and Leeuw 2002 for similar examples).

Many if not most policy processes are multi-actor in nature because the policy is formulated or negotiated in interaction and co-operation between different administrative bodies and layers, and often also with private and societal organizations. The same holds for many 'implementation' processes. In theses cases, what has been said about third actors above applies in a specific way. Like citizens, clients and media co-producers can influence evaluation and will react to their outcomes (or not) and thus co-determine the setup, conclusions and impact of evaluations. More specifically, in cases of co-production it is not a single actor who defines and commissions evaluations. Different actors may do so at the same time, thus producing concurring evaluations with potentially contradictory results, which will not necessarily contribute to substantive improvements or better cooperation. An interesting aspect in this connection is the question who initiates an evaluation, along which lines and especially: who is evaluated. Even apart from the substantive conclusion of an evaluation these elements may strongly influence sensemaking processes of the different actors and hence their reactions.

Dutch spatial developments actually are a co-production of different governmental bodies, private enterprises, environmental and other interest groups and citizens. If the central government or the Spatial Planning Bureau commissions an evaluation to assess to what extent the general goals formulated in the Spatial Policy Bill are realized, municipalities may experience this as an effort to increase central control, especially if they are not given the opportunity to insert their goals, limitations and assessments in the evaluation process (see Teisman et al. 2002, for a detailed analysis and some options for dealing with these complications).

This theoretical argument gives rise to the presupposition that the design, the outcome and the impact of evaluation are products of complex interaction processes in which sensemaking from a diversity of perspectives takes place. Then, insight in the relevant repertoires and knowledge of the interaction patterns between relevant actors helps to understand the dynamics around evaluations and their impacts. Based on such understanding ideas for more effective evaluations and evaluation arrangements may be developed.

How to evaluate public management reforms?

Above, I hypothesized that the relative lack of comprehensive evaluations of NPM reforms is due to the self-evidence that their rationales have for their promoters. Of course, from an 'ex-

ternal' perspective, such as developed in this paper, evaluation of NPM reforms as such is very desirable. In this section I develop some ideas on what such evaluations might look like. The following section is devoted to the question how evaluations of NPM innovations – and also evaluations within the framework of NPM practices - can be done in a way that enhances utilization of evaluation results and their impact. For both themes I make use of the theoretical argument developed before.

In de preceding discussion of the dynamics of evaluation I advanced the thesis that actual developments on the behavioral level (design of the evaluation, steps in the evaluation process, and (re)actions in relation to its outcomes) are mediated by sensemaking among actors involved in the consecutive phases. These sensemaking processes are connected to actor repertoires and their interaction patterns, hence they will be situation or position or sector specific.

This argument is not only relevant for evaluations of management reforms but may also be applied to the dynamics of the reforms themselves (cf. Teisman and Van Buren, chapter 11 of this volume). This is relevant in the present context, since it has consequences for the way in which evaluations of public management reforms can and should be undertaken. The argument runs as follows.

NPM reforms are designed on the basis of specific rationales (which could, in my language, be denoted as repertoires). Their actual shape and their impact on actual behavior, however, depends to a large extent on the sense other actors make of the reforms. This would imply that the meaning of such innovations may differ between actors and that their impact and 'success' is context dependent. To phrase it differently: the effects of an innovation do not only depend on characteristics of that innovation, but also – and primarily! – on the meaning it acquires from actors' perspectives and on the interplay of interpretation related actor behaviors. Thus, a typical evaluation question should not be whether new arrangement X is 'good', or 'better' than arrangement Y, but how arrangement X works in a specific context[5] and why (see also Pollitt et al 2004). Evaluations should indeed search for explanations of the functioning and effects of public management innovations since this may provide new input in sensemaking and learning processes, contributing to ideas for improvement. Moreover, the analytical argument can potentially be generalized to other settings, which is generally not the case for substantive findings. When a successful example of contract management is found (irrespective of the criteria that are used to arrive at that judgment), a key question should be what conditions and mechanisms produced that result. Insights like these are necessary to be able to link evaluation results to other settings.

[5] What the specific context is, is not trivial either. Actors may have different perceptions of or convictions about what the relevant context is.

The theoretical approach advanced in this paper also implies that knowledge of the perspectives (repertoires) of actors involved (in whatever role) as well as of their interactions is necessary for explanations to be found. Therefore, interviews with actors in different roles and positions – or other forms of their participation in the evaluation process – are required to reconstruct their repertoires and be able to explain their (re)actions and thereby the functioning and eventual effects of the management structure or strategy under review. It should be noted that this thesis is not based on a normative choice for democracy or empowerment (however honorable such a choice would be) but on an analytical argument on the dynamics of evaluation and public management innovations.

It should also be noted that involvement of stakeholders in the evaluation process may influence and enhance the impact of the evaluation since participants are more aware of the fact that there is an evaluation. Therefore, they may be more inclined to acquire knowledge on its content and conclusions, to recognize more elements, and to feel more committed to act upon the evaluation outcomes. In fact, resulting changes in actor sensemaking and behavior may also influence the functioning and effects of the public management innovation(s) under scrutiny. The direction of these changes depends on how evaluation results and other experiences in the evaluation process are linked by actors to their existing ideas and practices (Van der Meer 1999). That's why the idea that participation produces support seems to be too simplistic, both with respect to the evaluation process and with respect to the public management innovation evaluated. It is quite conceivable that a negative attitude towards a public management innovation is reinforced or induced in the process for some actors. And even if actors' attitudes change in a positive direction the question remains exactly what it is that thet support: their (changed) image of what the innovation is about is decisive.

Thus, although I have argued that some form of interaction with stakeholders in the evaluation is required for sensible evaluation of public management innovations, it should be realized that in doing so the object of investigation might change. Or, to put if differently, the evaluation process becomes an integral part of the functioning of the innovation.

Towards effective evaluation

The ideas on the dynamics of evaluation outlined above, also provide clues to improve the efficacy of evaluations, both of and in NPM innovations. The first question here is: how can we define and assess the effectiveness of evaluations? A number of levels of impact come to mind (cf. Hupe and Van der Meer 2002, 14):

1. actors evaluated and other stakeholders take note of the evaluation;
2. actors use evaluation results in considering changes in policy, management, working processes, etc.;

3. actors support the conclusions of an evaluation; and
4. actual changes in policy (implementation) or management as suggested by the evaluation.

However, on reflection this scheme is problematic.

First, although condition 1 seems to be valid on the level of the network as a whole – it is a necessary condition for any impact to materialize – it is not so at the level of individual actors. Through chains of interactions in the network the behavior of some actors may change, partly as a consequence of the evaluation, without these actors having any knowledge of the evaluation (indirect impact).

Second, level 4 may come about as a consequence of other processes than the evaluation. In fact it is generally very difficult to attribute actual changes to a single cause or even to estimate the contribution of a give cause.

Finally, levels 3 and 4 are not necessarily higher or better than level 2. Criteria, conclusion and recommendations given in an evaluation reflect, as I suggested above, in part a specific perspective or repertoire. Thus, evaluation results or other experiences in the evaluation process may acquire other meanings for other actors, making sense utilizing their own repertoire. Although the evaluators may judge this negatively, from any other stance it may be considered a potential contribution to new ideas, new innovations and hence learning.

Therefore, my thesis is that the only *general* criterion for effectiveness of evaluation is the extent to which it contributes to learning processes in the network of relevant actors/stakeholders. The extent to which this is the case can be related to the number or the diversity of actors learning (or reporting to learn) something from the evaluation, or to the measure to which new concepts or 'facts' from the evaluation are used in consecutive discourse. 'Learning' might also be measured by the extent to which actors attribute actual changes - or decisions not to change - to the evaluation or to concepts and 'facts' produced by the evaluation. Finally, learning can be measured by the extent to which evaluations trigger interaction and debate in the network.

One could point to the fact that many evaluations, especially in the context of NPM, do not primarily aim at learning, but rather at control and accountability (see Sanderson 2001, on this distinction; Teisman and Van der Meer 2002). However true that may be, my argument would be that the efficacy of accountability oriented evaluation depends on the extent to which and the way in which it is include in sensemaking processes of and among actors involved. If no actor reads something new in the evaluation, if its results are no object of debate and if no decisions (not) to change behavior can be related to the evaluation, the evaluation has not been effective.

Thus, the key question becomes how learning from evaluation can be enhanced (Cf. Thoenig 2003).

The theoretical argument in this paper suggests that at least some actors should both recognize the evaluation (outcomes) as sensible from their existing repertoire and as new. If the first condition is not met, the evaluation is simply noise to them, if the second is not met, there is nothing to be learned (except the fact that there is nothing to be learned, which consolidates existing repertoires and may decrease opportunities for future learning). The 'connectivity' of evaluations – the extent to which they are connectable to existing repertoires of different actors, thus inserting new element in these – depends on both their contents and the interaction processes in which they are embedded. Actors commissioning or performing evaluations can enhance connectivity by gaining knowledge of actor repertoires, related preoccupations, dominant questions etc., to which the evaluation design can be adapted. Alternatively they can involve second and third actors in the process of designing the evaluation.

It should be realized that the preconditions for connectivity and learning are generally different for different actors, since their repertoires and positions in interaction patterns differ. By consequence, it will not be possible to design evaluations that maximize learning for all actors at the same time. Actors have different interests and questions. Especially in the case of co-productive or participative policy processes, actors should be allowed, enabled and stimulated to perform or commission their own evaluations, thus feeding their own learning processes. This may also reduce perverse behavior or defensive reactions (however these may be defined).

But it should also be realized that policy and management results eventually are produced in the interplay between all actors. If evaluation is to stimulate and improve their cooperation, it should also help their mutual debate and collective learning processes (Gray and Jenkins 2003). Efforts should therefore be undertaken to enable linking of and confrontation between different evaluations, done by different actors at different levels, and to supplement them with joint (multi-actor) evaluations (Teisman et al. 2002; Van der Meer and Edelenbos 2006). In the latter case collective learning and mutual cooperation can already take place in the design phase of the evaluation.

Conclusions

In this paper I dealt with some issues concerning evaluation of and in NPM innovations. Evaluations in support of accountability constitute a core element of NPM philosophy and arrangements. Still they may have unintended and even perverse effects. It was also observed that NPM arrangements as such are far less systematically evaluated. To explain this state of affairs and to develop directions of improvement, I proposed a theoretical framework in which

'sensemaking', interaction and 'repertoires' are the core concepts. By conceiving of evaluation as a sensemaking process in interaction between actors in a network, both the design and the impact of evaluations can be explained (as well as the lack of evaluation or the lack of impact). Moreover the argument gave rise to some ideas about shaping evaluations of public management innovations as such. These can be summarized as follows:

- the success and impacts of public management innovations are not only determined by characteristics of the innovation; evaluations should therefore be contextualized;
- evaluations of public management innovations should include (context dependent) explanations for their functioning and effects, both for context related learning and for generalization to other contexts;
- these explanations require knowledge of actor repertoires, which implies some sort of participation of key actors and stakeholders in the evaluation process; and
- evaluation of NPM innovations influence the functioning and impact of these innovation, i.e. evaluation becomes part of the innovation.

Finally I derived some conclusions with respect to enhancing efficacy of evaluations (both of and in public management innovations):

- learning is the only general criterion for evaluation effectiveness;
- connectivity of evaluations in relation to actor repertoires determines their learning potential. Thus enhancing evaluation efficacy requires enhancing connectivity which in turn requires interaction with stakeholders; and
- since different actors have different repertoires and learning needs, a multiplicity of evaluations should be favored, performed by different actors, at different levels and with different core questions. However these evaluations should have enough connectivity to contribute to mutual debate and cooperation. Also multi-actor evaluations should be an element of the set of evaluations in a network.

Now, a focus on learning and actor involvement in the learning process is not new in the world of evaluation. On normative grounds or just to enhance utilization diverse forms of participative evaluations have been developed (Stake, 1983; Guba and Lincoln 1989; Cousins and Earl 1992; Abma 1996; Patton 1997; Preskill and Torres 1999; Van der Meer and Edelenbos 2006). However, the rationales of these forms of evaluation seem to have been seen as incongruent with the NPM paradigm for long. Still, gradually new connections between these different 'repertoires' seem to come into being, especially in situations in which the fact that policy outcomes result from the interplay between many actors cannot be missed (Teisman and Van der Meer 2002). For example, the authors in a volume edited by Gray et al. (2003) struggle with the role of evaluation in collaborative arrangements. On the one hand

there is a felt need for assessment of results of collaboration and of the success of collaborative arrangements by means of 'objective' and independent measuring, on the other hand the necessity to explicitly involve the actors in the network is underlined, although not all contributions are clear about what their roles should be. The framework proposed in the present chapter may help in this respect.

To be sure, further research is needed to test the theoretical notions advanced in this chapter, especially their application to NPM innovations and related evaluations. Also the recommendations should be tested and evaluated in their own right. This chapter intended to provide a framework and some first steps to perform such research and to arrive at evaluations that more effectively support the process of learning by experience on an organization and network level.

References

Abma, T. A. (1996). *Responsief evalueren: discourses, controversen en allianties in het postmoderne*. Rotterdam.

Bekkers, V. J. J. M. (1998). New forms of steering and the ambivalency of transparency. *Public administration in an information age: a handbook*. I. T. M. Snellen and W. B. H. J. v. d. Donk. Amsterdam etc., IOS Press**:** 341-357.

Boyne, G. A., C. Farrell, J. Law, M. Powell, R.M. Walker. (2003). *Evaluating public management reforms*. Buckingham/Philadelphia, Open University Press.

Broadbent, J. and R. Laughlin (1997). "Evaluating the 'new public management' reforms in the UK: a constitutional possibility." *Public Administration* 75: 487-507.

Christensen, T., P. Laegreid and L.R. Wise. (2003). Evaluating public management reforms in central government: Norway, Sweden and the United States of America. In H. Wollmann (ed.) *Evaluation in public-sector reform: concepts and practice in international perspective*. Cheltenham, Northampton (MA), Edward Elgar: 56-79.

Cousins, J. B. and L. M. Earl (1992). "The case for participatory evaluation." *Educational Evaluation and Policy Analysis* 14(4): 397-418.

Dawson, S. (1996). *Analysing Organisations*, McMillan.

De Bruijn, H., E. Ten Heuvelhof, and R. in 't Veld. (1998). *Procesmanagement: over procesontwerp en besluitvorming*. Schoonhoven, Acedimic Service.

Edelenbos, J. (2000). *Proces in vorm: procesbegeleiding van interactieve beleidsvorming over lokale ruimtelijke projecten*. Utrecht, Lemma.

Forbes, M. and L. E. Lynn Jr. (2005). "How does public management affect government performance? Findings from international research." *Journal of Public Administration Research and Theory* 15: 559-584.

Gray, A. and B. Jenkins (2003). Evaluation and collaborative government: lessons and challanges. *Collaboration in public services: the challenge for evaluation*. A. Gray, B. Jenkins, F. Leeuw and J. Mayne. New Brunswick (NJ), Transaction Publishers**:** 227-244.

Gray, A., B. Jenkins, F. Leeuw, J. Mayne (2003). Collaborative government and evaluation: the implications of a new policy instrument. *Collaboration in public services: the challenge for evaluation*. A. Gray, B. Jenkins, F. Leeuw and J. Mayne. New Brunswick (NJ), Transaction Publishers**:** 1-28.

Green, J. C. (1999). "The inequality of performance measurements." *Evaluation* **5**(2): 160-172.

Guba, E. G. and Y. S. Lincoln (1989). *Fourth generation evaluation*. Newbury Park, London, New Delhi, Sage.

Hazeu, C. A. (2000). *Institutionele economie: een optiek op organisatie- en sturingsvraagstukken*. Bussum, Coutinho.

Hupe, P. and F.-B. v. d. Meer (2002). Doorwerking van emancipatie-effectrapportages in beleidsprocessen. Den Haag, Ministerie van Sociale Zaken en Werkgelegenheid.

Kickert, W. J. M., Ed. (1997). *Public management and administrative reform in Western Europe*. Cheltenham, Northampton, Edward Elgar.

Meyer, M. W. and V. Gupta (1994). "The performance paradox." *Research in Organizational Behavior* 16: 309-369.

Patton, M. Q. (1997). *Utilization-focused evaluation*. Thousand Oaks, London, New Delhi, Sage.

Pollitt, C. (1995). "Justification by works or by faith? evaluating the new public management." *Evaluation* 1(2): 133-154.

Pollitt, C. and G. Bouckaert (2000). Public management reform: a comparative analysis. Oxford, Oxford University Press.

Pollitt, C. and G. Bouckaert (2003). Evaluation in public sector reforms: an international perspective. *Evaluation in public sector reform*. H. Wollmann. Cheltenham, Northampton (MA), Edward Elgar: 12-35.

Pollitt, C., X. Girre, J. Lonsdale, R. Mul, H. Summa and M. Waerness (1999). Performance or compliance? performance audit and public management in five countries. Oxford, Oxford University Press.

Pollitt, C., C. Talbot, J. Caulfield and A. Smullen (2004). *Agencies: how governments do things through semi-autonomous organizations*. Basingtoke, Palgrave/Macmillan.

Power, M. (1994). *The audit explosion.*`. London, Demos.

Preskill, H. and R. T. Torres (1999). *Evaluative inquiry for learning in organizations*. Thousand Oaks, London, New Delhi, SAGE.

Ryan, K. and L. DeStefano (2001). "Dialogue as a democratizing evaluation method." *Evaluation* **7**(2): 188-203.

Sanderson, I. (2001). "Performance management, evaluation and learning in 'modern' local government." *Public Administration* 79(2): 297-313.

Schein, E. (1996). "Drie managementculturen: de sleutel tot bedrijfsleerprocessen." *Holland/Belgium Management Review* **51**: 25-35.

Stake, R. E. (1983). Program evaluation, particularly responsive evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (red.), *Evaluation models*, Boston: Kluwer-Nijhoff, pp. 287–310.

Taket, A. and L. White (1997). "Working with heterogeneity: a pluralist strategy for evaluation." *Systems research and behavioral science* **14**(2): 101-111.

Teisman, G. R. and F.-B. Van der Meer (2002). Evalueren om te leren: naar een evaluatiearrangement voor de Vijfde Nota RO. Rotterdam, Erasmus Universiteit.

Teisman, G. R. and T. J. M. Verheij (1996). Draagvlakvorming bij technisch-complexe projecten. *Grote projecten*. J. A. d. Bruijn, P. d. Jong, A. F. A. Korsten and W.P.C. van Zanten. Alphen aan den Rijn, Samson H.D. Tjeenk Willink: 174-192.

Thoenig, J.-C. (2003). Learning from evaluation practice: the case of public-sector reforms. *Evaluation in public-sector reform: concepts and practice in international perspective*. H. Wollmann. Cheltenham, Northampton, Edward Elgar**:** 209-230.

Tops, P. W. and R. Weterings (1998). Gemeentelijk beleid en co-productie. *Lokaal bestuur in Nederland, inleiding in de gemeentekunde*. A. F. A. Korsten and P. W. Tops. Alphen aan den Rijn, Samsom: 518 - 528.

Van der Meer, F. B. (1999). "Evaluation and the social construction of impacts." *Evaluation* 5(4): 387-406.

Van der Meer, F.-B. and J. Edelenbos (to be published 2006). "Evaluation in multi-actor policy processes: accountability, learning and cooperation." *Evaluation*.

Van der Meer, F. B., G. J. D. De Vries and G. A. N. Vissers (2000). "Evaluatie en leerprocessen bij de overheid: de rol van institutionele condities." *Beleidswetenschappen* 14(3): 253-277.

Van Thiel, S. and F. Leeuw (2002). "The performance paradox in the public sector." *Public Performance and Management Review* **25**(3): 267-281.

Weick, K. E. (1979). *Social psychology of organizing*. Reading (Mass.), Addison-Wesley.

Weick, K. E. (1995). *Sensemaking in organizations*. London, Sage.

Weiss, C. H. (1990). Evaluation for decisions: is anybody there? does anybody care? *Debates on evaluation*. M. c. Alkin. Newbury Park, London, New Delhi, Sage**:** 171-184.

Wollmann, H. Ed. (2003). *Evaluation in public-sector reform: concepts and practice in international perspective*. Cheltenham, Northampton (MA), Edward Elgar.