SJPA
16(3)

# Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable

Peter Dahler-Larsen*

*Peter Dahler-Larsen*
Department of Political Science, University of Copenhagen

**Abstract**

Evaluability assessment is a diagnostic and prescriptive tool which helps evaluators determine whether evaluation is appropriate in a given situation. Thus, evaluation is understood as a situational good. Today, however, evaluability assessment is no longer particularly popular. Mandatory, comprehensive and repetitive evaluation systems are gaining ground in public administration supported by general social, political and managerial norms and values, indicating that evaluation is believed to be a universal good. Can a form of evaluability assessment be re-vitalized in order to pave the way for a more modest, more reflexive, and more context-sensitive belief in evaluation? The article offers a specific list of items in an updated version of evaluability assessment, and concludes with a discussion of the limitations of such approach.

Keywords: Evaluability assessment, evaluation, evaluation system, situational good, universal good

**Evaluering som et situationelt eller et universelt gode?**
**Hvorfor evaluability assesment for evalueringssystemer er en god ide, hvordan den kunne se ud i praksis, og hvorfor den ikke er på mode**

Evaluability assessment - "evaluerbarhedsvurdering" – er et diagnostisk og præskriptivt redskab, som hjælper evaluatorer med at afgøre, om evaluering er egnet i en given situation. Således er evaluering et situationelt gode. I dag er evaluability assessment imidlertid ikke længere særlig populær. Obligatoriske, omfattende og repetitive evalueringssystemer vinder frem i offentlig forvaltning støttet af generelle sociale, politiske og ledelsesmæssige normer og værdier, hvilket er tegn på en tro på evaluering som et generelt gode. Kan en form for evaluability assessment genoplives for at bane vejen for en mere beskeden, mere refleksiv og mere kontekst-sensitiv tro på evaluering? Artiklen tilbyder en konkret liste over komponenter i en sådan opdateret version af evaluability assessment. Der konkluderes med en diskussion af begrænsninger i en sådan tilgang.

Nøgleord: Evaluerbarhedsvurdering, evaluering, evalueringssystem, situationelt gode, universelt gode.

Offentlig förvaltning
Scandinavian Journal of Public Administration
16(3): 29-46

ISSN: 1402-8700
e-ISSN: 2001-3310

*__Peter Dahler-Larsen__, PhD, is professor at the Department of Political Science, University of Copenhagen. His research interests focus on cultural, institutional and political aspects of evaluation. As a constructivist, he is especially interested in how evaluation is constructed and in how the constitutive effects of evaluation are produced. He was president of European Evaluation Society 2006-07. One of his most recent publications is *The Evaluation Society* (Stanford University Press 2012)

## Introduction

How can one think intellectually about the role of an evaluation system before it is implemented? The issue at stake is not only how to design such systems, but also prior to that to distinguish between situations where an evaluation system is a very good idea and situations where it is not.

The field of evaluation has developed a heuristic tool called "evaluability assessment" (EA) which is supposed to help evaluators make a situational diagnosis. EA incorporates a rationalistic view of knowledge according to which each piece of knowledge should be bought stepwise and rational decisions should be made at each step.

In that view some activities, programs and policies are not yet "ready" for evaluation, and they should be "straightened out" before evaluation can proceed. Evaluation is appropriate in some situations, but not in all, says EA. It is an important distinction, theoretically, culturally and normatively, whether one believes that evaluation is a universal good regardless of situation and context, or whether it is merely a situational good that should be applied only where it is fit. EA is an expression of the latter of these beliefs.

The idea of EA was popular in the 1970s, but the idea largely died out (Smith, 2005). It might have been based on too rationalistic and simplistic assumptions about how evaluative knowledge is used. It was designed to be applied to stand-alone evaluations, not evaluation systems. It may have been made obsolete by today´s belief in evaluation as a *universal good* so that the perceived need to check whether evaluation is appropriate under particular circumstances disappears.

This view is consistent with Vedung´s (2010) view of four waves in the history of evaluation: the science-driven wave, the dialogue-oriented wave, the neoliberal wave, and the evidence wave. While the two first waves typically consisted of evaluations crafted one by one, the two latter waves prescribe the construction of evaluation *systems* (Leeuw & Furubo, 2008), such as those known under the names of performance management, auditing, accreditation, indicators, and testing regimes.

When these forms of evaluation and monitoring can be regarded as systems, it is because they become permanent, routinized, and extend across time and space. This change signifies a greater belief in evaluation as a universal good. It is the *capacity* to evaluate that is good (Baizerman, Compton & Stockdill, 2005), it is an evaluation *culture* that induces change, and evaluation *policies* (Trochim, 2009) and evaluation *systems* which bring order and efficiency. Needless to say, paraphernalia related to evaluation systems constitute an important segment of what is today accepted as effective and legitimate tools of public government and administration. Under these circumstances it is less important for evaluation to demonstrate the likelihood that the actual benefits of evaluation in a situation at hand will outweigh the costs. It has not always been so. It was not so when EA was popular.

The shifting balance between evaluation as a situational good versus evaluation as a universal good is the overall theme of this paper, and more specifically, of course, *whether a reinvention of EA might help reinstall the view that the benefits of evaluation in a given situation should be justified rather than just assumed.*

In other words, to borrow a metaphor for generations of software, is it possible to move from early EA 1.0 to contemporary EA 2.0, so that today´s needs concerning the assessment of evaluation systems can be better met?

The strategy of the article is the following: First, evaluation systems and the challenges they bring are described. Next, the paper goes back to early evaluability assessment (EA 1.0). The purpose of EA is to establish situational criteria determining when evaluation is appropriate and when it is not. Next, it is suggested what EA would look like if updated to our era of evaluation *systems*. The practical relevance for public administration of this move is obvious. If evaluation systems are more appropriate in some situations than others, these different situations should be identified and consequences should be drawn. Finally, it is discussed whether EA 2.0 (for evaluation systems) is likely to be adopted in practice. Perhaps the idea is good, but not in fashion.

## Evaluation systems

There has been a paradigm shift in the organization of evaluation. The classical paradigm for evaluation was an in-depth, expert-based, ad-hoc inquiry, but today, more emphasis is on evaluation systems (Leeuw & Furubo, 2008).

Evaluation systems are diverse in form, shape and maturity, and thus not easy to define. But they have the following ideal-typical characteristics. They are fairly permanent, repetitive, and routine-based. They generalize forms of evaluation and/or their results across time and space. Evaluation systems are decreasingly dependent on the values and ideas and styles of individual evaluators. Instead, they embody evaluation epistemologies or institutionalized types of thinking, and they are supported by general and abstract tools such as verification processes, documentation processes, indicators, criteria, standards, benchmarks, testing systems, information technology and handbooks that can be used in fairly standardized ways across different substantial areas of activity. Evaluation systems allow the handling of information about large amounts of public activities in a systematic, integrated, and comparable way.

Evaluation systems are embedded in organizational procedures of verification and undergirded by organizational responsibilities. Evaluation systems are run by organizations. Evaluation systems produce streams of evaluative information (Rist & Stame, 2006) rather than stand-alone evaluation reports. Evaluation systems include systems of performance management, systems of audit, inspection and oversight, accreditation systems, and monitoring systems (Leeuw & Furubo, 2008). In other words, I include quite a number of diverse practices under the rubric “evaluation systems”, if only they have the characteristics described above. I do this because I believe my argument about evaluability assessment is applicable to everything that has evaluation system characteristics.

## Evaluation systems as a social and political phenomenon

The emergence of evaluation systems is probably due to a large and complex configuration of factors which are both symbolic and functional, such as the following:

After years of debates with ad-hoc evaluations with failing utilization, there has been an increasing need to develop evaluation capacity in organizations (Baizerman, Compton & Stockdill, 2005), to enhance evaluation cultures, and to create systematic managerial and organizational approaches to ongoing evaluation so that evaluation is better integrated and mainstreamed into organizational processes. Stand-alone evaluations often had little impact. In that sense, evaluation systems are a meaningful response to the most classical issue in the field of evaluation, ie. the failing utilization of evaluation.

In addition, many ad-hoc evaluations were based on a broad variety of evaluation models corresponding to a kaleidoscope of different social, cultural and paradigmatic perspectives, but these many viewpoints did not add up to a new and coherent social agenda (Boltanski & Chiapello, 2007). Evaluation processes have often been unpredictable and it has been difficult to synthesize evaluative knowledge into a managerial or steering perspective without more integrated approaches. All this paved the way for evaluation systems.

Through this abstraction – evaluation systems – complexity is reduced considerably. Evaluators no longer need substantial insight in what is evaluated, but can rely on broad assumptions about the virtues of particular organizational recipes (Røvik, 1998; Meyer, Boli & Thomas, 1994). The abstraction from "things done" to "systems controlling how things are done" is also highly beneficial for the evaluator/inspector if his/her expertise is methodological, general and abstract rather than connected to a given substance area.

The institutional advantages (or potential disadvantages) of evaluation systems depend to a large extend on demonstrability, auditability and verifiability (Power, 1996: 302). Evaluation systems must be in place in organizations in order to render organizations auditable, evaluable, inspectable, and certifiable. The primary function of an evaluation system may not be to monitor quality, but to guarantee internal or external auditability (Power, 1996: 300).

Perhaps the interest in risk avoidance and risk management is further enhanced by a society which since 2001 has been occupied with monitoring and surveillance as a medicine against terrible, catastrophic events which in fact rarely occur (Dahler-Larsen, 2012).

The evaluation industry has not hesitated to exploit the opportunities which this situation offered. The market for evaluation culture, evaluation capacity, evaluation policies, and evaluation systems may be larger and more rewarding than the market for bare evaluation, whether or not the prizes to be won is profit or institutional power. It has been possible for the evaluation industry to expand its market exactly by moving from singular evaluations to a focus on evaluation systems (Power, 2005).

In Power's (1996) analysis, the self-gratulatory process of evaluation systems checking evaluation information fits into a larger social project of "producing comfort." Hood (2002) argues that risk is managed and blame is shifted, as politicians seek to install "quality assurance systems" which – in the name of accountability – often tends to be used as risk-placing, blame-placing and responsibility-avoiding mechanisms by politicians themselves. With the intense media focus on potential scandals, the motivation of politicians to install self-protecting mechanisms is only further enhanced. "More monitoring of various kinds is an easy and politically acceptable solution to perceived problems and scandals", says Power (2005: 341). In this way, a cultural mentality in support of quality assurance and evaluation systems is supplemented with political power. Evaluation systems are indicative of larger set of beliefs in evaluation in contemporary society or what Schwandt (2009) calls an evaluation imaginary.

Evaluation systems can not only perform an information function, but also a resource allocation function, a legal function, and a political function. They are to an increasing extent supported by regulatory institutional pillars (Scott, 1995).

In this light, an increasing amount of literature suggests that evaluation systems may have a number of negative and perhaps unintended effects, including that they enhance single-loop learning but hinder double-loop learning, that they provide only procedural assurance, that they focus on performance but not on the assumptions undergirding existing policies (Leeuw & Furubo, 2008: 165), that they incur large hidden costs (Power, 2005: 335)**,** that they are marred by a performance paradox so that more measurement does not lead to more quality (van Thiel & Leeuw, 2002), and that evaluation systems have constitutive effects on practice. Constitutive effects refer to the ways in which evaluation systems shape behaviour and redefine the meaning of public activities because evaluation indicators become goals in themselves (Dahler-Larsen, 2012). Although evaluation systems may increase transparency and enhance manageability, they may thus incur a new set of risks, some of which are transplanted down through the political and administrative chain of command, where others than those who installed these systems then seek to avoid the risks associated with measurement problems and potential low scores (Rothstein, Huber & Gaskell 2006).

Space allows neither a further elaboration of causes of the development of system, nor a further inquiry into the normative, theoretical and empirical foundations for the critique of evaluation systems. Instead, one particular question shall be taken up. The question is whether it can be determined in advance if an evaluation system is more or less appropriate in a given situation. To answer that question, we shall consult an idea that was developed for the same purpose in relation to evaluation, not evaluation systems, and find out whether that idea can be updated and be made relevant for our time, the era of evaluation systems. That idea is called evaluability assessment (EA).

## Evaluability Assessment 1.0

Evaluability assessment is a process which leads to a decision about whether it is sensible to evaluate under given circumstances (Wholey, 1987; 2004). The idea is practical and useful if one wants prescriptions about when to evaluate and when not to. The main question in EA is not whether evaluation *can* be done, but *whether it is a rational thing to do* under the circumstances in the light of the expected improvements coming out of the evaluation (Shadish, Cook & Leviton, 1991: 237). [1]

In EA, circumstances in an evaluand and its context are clarified before evaluation is undertaken. *One potential outcome* of EA is that problems with the program that makes it non-evaluable are straightened out (much in the same way that a hair-dresser combs your hair before cutting it) before evaluation can proceed. In this sense, EA already serves an early formative evaluation function. But when the "problems" of the program have not been straightened out yet or perhaps cannot be straightened out in the near future, or perhaps will never be straightened out, then *another potential outcome* of EA is that evaluation is deemed not appropriate for the situation. Resources may then be spent better on evaluating other things or on other things than evaluation. [2]

In EA 1.0 the following specific questions should be answered (Shadish, Cook & Leviton, 1991: 237; Rossi, Freeman & Lipsey, 2004: 137):

**a.** Is there a clear description of the program? If not, resources are better spent on clarifying the program rather than on evaluating an unclear one.

**b.** Has the program overcome known implementation problems? The ability to draw clear conclusion about the program is improved dramatically if implementations problems are removed so they no longer can be responsible for program failure. Implementation problems can be known or unknown. It is evidently acceptable to use evaluation to identify the scope and nature of unknown evaluation problems, so that is not the issue here. Our criterion is about known implementation problems. Known implementation problems do not require extraordinary evaluation. Managers and others should sort out already-known implementations problems as soon as possible and not wait for evaluation. If consensus can be reached about what changes in program design are appropriate, these changes can be enforced without the costs of a large-scale evaluation (Wholey, 2004). Thus, the criterion reminds us that evaluation systems should not be functionally overburdened with problems that ought to be handled through normal managerial operations in the organization.

**c.** Is there a fairly good program theory, ie. a logical explanation of why the intervention should lead to expected outcomes? If not, it is better to clarify the logic of the program and perhaps improve it accordingly before evaluation is undertaken.

**d.** Are there well-described, plausible, and realistic goals? If not, the findings of the evaluation are predictable even without evaluation.

**e.** Are relevant data within reach? If not, evaluation resources could be spent better on alternatives to evaluation.

**f.** Are opportunities to improve the program identified? If intended users of the evaluation are not able or willing to use the evaluation results, the evaluation is less likely to make a practical difference.

## What happened to evaluability assessment

Attempts have been made to invigorate EA since its golden age in the 1970s and it is still promoted by some, for example for educational purposes (Leviton et al., 1998). Others suggest that although the interest in EA is not growing in evaluation, it may have a new life in some disciplines (Trevisan, 2007).

However, Smith correctly concludes that, *the idea is not nearly as popular today as it was in the 70´ies* (Smith, 2005: 137). Although, of course, it may take place here and there as part of evaluation practices that are rarely documented and described, EA does not seem to play an important role in the official rhetoric and discourse about evaluation[3]. In fact, the birth and decline of EA is a very significant development in the history of evaluation. The field as such has – with the help of a number of external factors in the surrounding society – managed to get around the disturbing possibility that in some situations, evaluation would not be welcome.

Several reasons may help explain this. For example, EA as a concept is not sufficiently articulate and there is a lack of a clear EA methodology (Trevisan, 2007: 291). However, to the extent that good evaluation is a result of situational judgment and wisdom, the spirit of EA should thrive even if algorithms for doing it may change a bit over the years.

Furthermore, it may be difficult to distinguish between EA on the one hand and pre-evaluation and formative evaluation activities on the other, for which reason it may be less fruitful to maintain an idea that EA is a distinct activity (with its own name and its own literature) separate from evaluation as such.

Next, EA as defined above may be appropriate only if the intended subsequent evaluation is an old-school goal-based evaluation focussing on whether clear goals have been achieved through proper means-ends relations. Since the birth of EA in the 1970s, however, a variety of evaluation models have emerged, including transformative, participatory and constructivist evaluations that do not require clear, consistent and agreed-upon goals, but proceed under the assumption that interests, perspectives and goals of various stakeholders are emerging, conflicting and can be dealt with during the evaluation itself.  On the other hand, the breaking-up of EA into separate tasks of which only some are dealt with in particular evaluation models may lead to loss of that type of overview which EA could provide (Smith, 2005: 139).

EA rests on an assumption that program development, implementation, EA, and evaluation are organized in an orderly fashion and that the reasons to move from one phase to the next are motivated by rational decisions. However, perhaps precisely due to these overgrown assumptions of rationality, EA has not gained ground since the 70´ies (Smith, 2005: 139). EA has gone down as evaluation systems have gone up, more or less.

The ideology of evaluation may have moved from regarding evaluation as an instrumental situational good to a universal good. It is the belief in the long-term principle of evaluation that counts. In an era of evaluation culture, - capacity and -systems, a singular evaluation does no longer have to demonstrate its rational benefits.

However, universal goods and situational goods are relative concepts, not absolutes. Even large-scale evaluation systems that cover a large ground in time and space are, of course, installed in a particular situation. Let us, for the sake of the argument, develop a version of EA as it would look if it were to support rational decisions about evaluation systems.

## Evaluability Assessment updated

A few of the items in this version of EA (presumptuously called EA 2.0 in brief) are fairly pedestrian pieces of advice, or repetitions of good advice offered in other works (Fitzpatrick, Sanders & Worthen, 2004; Wholey, 2004), but there are also a few points which are not trivial.

The items are organized as a list of factors which deserve to be considered. Although a stepwise algorithm, where a “no” to item one cancels all other items further down the list (see eg. Fitzpatrick, Sanders & Worthen, 2004) is logically attractive, I have not copied such approach, because an all-things-considered-approach has some critical advantages.[4] I offer the following list as a net list of factors for consideration. It is up to practical judgment in a given situation to determine the relevance of each factor – before an evaluation system is put in place.

### Characteristics of the evaluand

**1**. Does the object of evaluation have enough social impact or importance to warrant a formal evaluation system? (Fitzpatrick, Sanders & Worthen, 2004: 186). The underlying norm behind this question is that if society has limited resources for evaluation, they should be spent on the most important issues. Research on evaluation, however, suggests that society´s evaluation focus is sometimes very selective. Social work is often evaluated. War is not. Other examples of phenomena rarely evaluated include tax systems, courts, royal families, inspection, and public management ideologies. For example, New Public Management initiatives have not been evaluated nearly as much as they deserve (Pollitt, 1995).

**2**. Are the characteristics of the evaluated activities of such a substantial nature that they are appropriately represented by the indicators, standards and criteria of the evaluation system? An answer to this question implies an attention to the task structure at hand and to the nature of the activities, eg. their substantial diversity, and whether they are one-sided or two-sided (Abma & Nordegraaaf, 2003). The latter dimension refers to whether the user plays a substantial role in producing a successful outcome of the public service. Two-sided activities are for example therapy, learning (not teaching!), and prevention of risky health behaviour. Criteria that look at delivery of public service only at the supply side do not sufficiently capture the nature of two-sided public services.

Although objects of evaluation may be complex, objects of evaluation systems are likely to be even more complex, because they comprise not only specific interventions, programs or policies, but often whole sectors or areas of activity such as "schools". In other words, the risk of a reductionist view of the evaluand is higher for evaluation systems than for evaluations. An attention to the diversity and complexity of activities under evaluation is an important aspect of determining the situational appropriateness of a given evaluation system, depending of course, on the ability of that evaluation system to reflect such complexity.

**3**. How clear are the goals of the evaluated activities? "Clarity" is not only a matter of articulation, but also of the political landscape around policy-making. Is there agreement about the goals of, say, schools and universities? Only if goals are clear and consensual can an evaluation system be based on operational evaluation criteria that fairly represent these political goals. More often than not, there is some discrepancy between the two.

Makers of evaluation systems often become de facto policy makers (Power, 2005: 335) because evaluation criteria are constructions which are not direct representations of political goals.

Even if an evaluation system is politically sanctioned and thereby legitimate, it does not logically follow that its criteria are also representative of already-legitimate political goals. Then, consistent with the active view of knowledge presented in an earlier section, the criteria inherent in an evaluation are not only descriptors of reality, but also active players in the socially valid definition of goals. Especially in a contested political environment, an attention to this definitory and constitutive rather than merely descriptive aspect of goal-setting is an important aspect of an evaluation system. (We shall come back to that under the heading of "actual consequences of evaluation systems".)

**4**. Does the problem structure of the evaluated activities warrant an integrated evaluation system? If the area of activity under evaluation is a response to a diverse set of complex problems, is the best approach then an integrated evaluation system? Or would it be more appropriate to tackle some of these uneven problems with different evaluation approaches each designed according to the nature of the problem?

**5**. Does the accountability structure of a practical field in which activities take place warrant an integrated evaluation system? By accountability structure I mean how a more or less clear definition of accountability leads someone to report to someone else about how some activity has been carried out (Pollitt, 2010). Formal accountability structures are often linear and hierarchical. Each unit is held responsible for how well-defined processes are carried out or for the changes in a small set of indicators.

In other situations, problems are complex, they require complex social coordination, and the social responsibility for their solution cannot be or is not pinned down to atomistic units in an administrative structure. Any evaluation system reflects an explicit or implicit vision of accountability, but how well does this vision match the accountability structure characterizing the practical field in which evaluated activities take place? Do evaluation systems reinforce the “silo problem” in public administration by overemphasizing micro-accountability for large social problems so that a broader cooperative effort is undermined?

**6**. What is the knowledge structure and theory structure related to the evaluated activities? In areas where there is already a well-developed knowledge base about professional activities, how well does the evaluation system take that into account? If there is a need to challenge existing theories, how well equipped is the evaluation system to produce a solid theory-based evaluation that does so? (Leeuw & Furubo, 2008). The negative scenario is one in which the evaluation system moves forward with institutional power, but without expertise and insight.

## Alternative knowledge streams

**7**. How well does the evaluation system take into account alternative, competing and supplementary streams of knowledge so characteristic of the knowledge society? Instead of assuming that without the evaluation system, there would just be ignorance (an old-fashioned assumption), it may be safer to assume that the situation in the knowledge society is characterized by large amounts of information in uncoordinated streams. An evaluation system does not offer an alternative to no knowledge at all, but is in fact in competition or in cooperation with many other forms of knowledge.

Does an evaluation system just require already-existing knowledge to be collected and documented one more time? How many agencies around a public institution should collect this information? How many quality centres, accreditation centres, evaluation institutions, consulting companies, prognosis-makers, statistical offices, think tanks and audit offices should a public institution report to? If the information provided by an evaluation system is really non-redundant in relation to other such systems, how well are the streams of information coordinated? Is added-value produced in the interaction between these streams of knowledge? If each evaluative organization argues that the information it collects is unique and necessary, a whole set of evaluative organizations may collectively overproduce knowledge that is under-coordinated. Is this problem thought-through before an additional evaluation system is installed?

## The characteristics of the evaluation system

**8**. If a techno-structure is necessary to implement the evaluation system, how well implemented and how reliable is that techno-structure? An entertaining negative example (for observers, not for participants) is the development of a national Danish testing system in schools. More than once did teachers and pupils prepare for the national computerized test, and subsequently Danish citizens read in the news about stalled computers, black screens, cancelled tests and frustrated pupils, teachers, and principals. The company delivering the digital infrastructure had not succeeded in developing a system that worked reliably in the national scale. EA would suggest that the evaluation should be made functional and then national, not national and then functional, as it happened.

**9**. How is the evaluation system anchored in an organizational structure which can infuse the system with expertise, support and legitimacy? Evaluation systems located in different organizational structures have different strengths and weaknesses. These locations differ with respect to broad social legitimacy, specific expertise, and evaluation capacity, all of which should be understood in relation to the evaluation of specific evaluands. For example, the Danish Evaluation Institute in education was established "from scratch", which made it fairly independent, but also rendered it disconnected from research expertise in education and in evaluation. In Sweden, tests in schools are developed by sharp academic institutions, in Denmark by a consulting company. Some SAIs (Supreme Audit Institutions) have a good social reputation, but may run the risk of losing that reputation if they perform forms of evaluation that deviate a lot from classical audit, on which SAIs may have an institutional monopoly, great power, and sufficient expertise. SAIs cannot count on the same privileges when they perform a broader variety of evaluations in addition to classical audit. Then knowledge contributed by the SAI may be one among many forms of knowledge, contested, debated, and criticized along many criteria such as relevance, usefulness, meaningfulness, and appropriateness, as sometimes happens in the knowledge society.

**10**. Is the evaluation system able to provide reliable, trustworthy information? (Fitzpatrick, Sanders & Worthen, 2004: 186). This question covers whether there are incentives to manipulate or misrepresent data and whether the evaluation systems has sufficient integrity to protect, analyze and report data, etc.

**11**. Are the costs of the evaluation system well described? And can a well-functioning evaluation system be built for that amount? Evaluation systems are not very good at measuring their own costs. The costs of evaluation systems do in fact include not only direct financial costs, but also the time professionals and others need to take out of their daily work time in order to feed the evaluation system with documentation (Power, 2005: 335). This amount of time can sometimes be reduced by integrating the documentation necessary for evaluation directly into work practices in intelligent ways, eg. through computerization. Still, the introduction of an evaluation system is often not based on a fairly exact cost-benefit analysis.

Such calculation is very much in the spirit of EA, ie. the calculation of likely benefits versus likely costs of designing, installing, and running the evaluation system. Often the argument for an evaluation system is that there is a need for the system (such as the need to improve quality) or that there will be some benefits (quality will be improved), or transparency is a goal in itself, but the costs of evaluation systems are often ignored, so the ideological calculus is always positive. EA suggests to make a cost-benefit analysis of the evaluation system, even in rough terms, and only to introduce evaluation systems where the analysis suggests that the benefits outweigh the costs, and where a functional evaluation system can in fact be built for the resources allocated to that purpose. An implication is that it is rational to live with minor local inefficiencies if they are not more costly than a general evaluation system which pinpoints them.

**12**. Is the evaluation system infused with an overarching ideology that tends to make the evaluation system self-justifying, or does the ideology of that evaluation system match the actual characteristics of the evaluation system and the context in which it operates? Is there a well-argued set of assumptions justifying how the design of the evaluation is intended to produce particular effects, for example based on a typology of forms of such systems (Foss Hansen, 2011), or is the evaluation system merely a political result of an "expectation gap" in society, where the public demands more comfort than can actually be delivered? (Power, 1997)

### The likely use and consequences of the evaluation system

**13**. Are there real opportunities for stakeholders to act in such a way that the intended use of the evaluation system can be fulfilled?

It is nice if there is agreement among central stakeholders about the intended use of the evaluation system, and some use it as a criterion in EA (Fitzpatrick et al., 2004: 186). But in a complex world, where the use of knowledge does not always match what is predicted, consensus is not a guarantee that the evaluation will actually function as promised. And consensus may not be easy to achieve, and the evaluation system will work without it.

In fact, the users of evaluation systems may be dispersed in different roles and positions inside and outside of organizational systems (such as managers, professionals, clients, and politicians). To overcome this complexity, it is often tempting to claim that the official purpose of an evaluation system is "learning" or "improvement of quality" because these purposes often have broad and positive connotations. However, unless "quality", "learning" and "improvement" are more specifically defined, and unless the evaluation systems is actually connected to learning opportunities and learning fora in organizations, the discrepancy between the official purpose of the evaluation and its actual use as perceived by a variety of stakeholders may be striking.

It has always been good advice in EA to check whether specific decision makers are in position to use the evaluation results, but the "use" is often more complex in the case of evaluation systems because of the diversity of stakeholders and because knowledge may be used in complex, dynamic, and non-linear ways.

If there is not agreement among key stakeholders about the intended use of the evaluation system or if consensus about broad positive intended uses is of little value, is it then possible to focus on fewer stakeholders who can use the information effectively? (Fitzpatrick et al., 2004: 186). Or will continued disagreement between various stakeholders determine the actual use of the evaluation system?

If politicians are intended to be a key group of stakeholders using the evaluation system, how consistent is that intention with what we know about how politicians already use such knowledge?

Another especially interesting group of stakeholders in the knowledge society is users of public services who find evaluative information on the internet. To what extend has the demands of such users and their actual patterns of use of information been understood before it is claimed that evaluative information must be made publicly available? These questions about the likely actual use of evaluative information by politicians as well as citizens are extremely relevant because there is very limited body of research that documents it (Pollitt, 2006). There may be good democratic reasons for publishing evaluative data, but if the argument in favour of publication is a specific *use* argument, assumptions about users and their demands and their behaviour should be an integrated part of the justification for an evaluation system.

If a specific category of stakeholders is pinpointed as crucial users of evaluation system, is a large segment of their decisions dependent on the information that the evaluation system provides, or more likely to be influenced by other factors? (Fitzpatrick et al., 2004: 186). If the latter is the case, evaluation systems may make little difference.

**14**. Has the evaluation system been piloted so that it has demonstrated some positive effects in practice and so that evaluation system can be improved based on actual experiences? When evaluation systems are introduced in complex organizational settings, it is often necessary to develop the design of the evaluation system iteratively in interaction with reality. If the motivation behind the evaluation system is a political desire to control and manage risk, a mandatory system here and now may be the answer. Without piloting, however, it is difficult to predict if the evaluation system may be technically dysfunctional, may meet unforeseen organizational resistance, or may have unforeseen negative consequences. Since evaluation systems are fairly permanent, comprehensive and often mandatory, their consequences may be of a much larger scale than stand-alone evaluations.

The national testing system in Denmark mentioned above is an example of an evaluation system which would have benefitted from pilot testing.

**15**. Have the consequences of the evaluation system (apart from its official purpose) been investigated? In an EA 2.0 perspective, it is beneficial to ask: How are people under the evaluation system likely to behave if they take the evaluation criteria seriously? In other words, does "if the activity is good, evaluation criteria will be met" also mean that "if evaluation criteria are met, the activity is good" (Munro, 2004: 1086)? If no, this indicates that the evaluation system may produce uncomfortable constitutive effects (Dahler-Larsen 2012).

Next, are initiatives such as meta-evaluation planned or implemented so that the actual consequences of the evaluation system can be checked once it is in operation? Are observations about constitutive effects taken seriously? And are the actual consequences seen in a broad perspective so that it includes whether or not evaluation systems have positive motivational effects on professionals, and whether evaluation systems leads to social trust in professionals, whether the risk-avoidance which motivated the evaluation system in fact creates new risks and pushes risk and blame around in society? (Hood, 2002; Rothstein, Huber & Gaskell, 2006)

**16**. Have alternatives to evaluation been considered? Does an analysis of a broad set of factors influencing decisions about the quality of particular services (such as education, organizational cultures, management structures, incentives, HR, and professional ethics) suggest that evaluation is *the* most productive way to better quality?

### Democratic aspects

**17**. How mandatory does the system have to be?  If there are benefits of some evaluation systems it does not logically follow that there are benefits from all mandatory evaluation systems, too. True enough, on the one hand, organizations which have severe quality problems may be organizations who are least likely to evaluate on a voluntary basis. On the other hand, the effects of a new organizational recipe (such as evaluation) may be more limited among organizations which are forced to adopt it than among organizations who adopt it voluntarily (Scott, 1987). Although, of course, some evaluators begin with an unquestioned legal requirement for evaluation,  the mandatory character of evaluation systems should not be regarded as a constant, but as a variable that can be controlled intelligently.

**18**. Are the democratic aspects of the evaluation system at hand thought through? By democratic aspects I here refer to the capacity of society to regulate its own functioning in a rational and autonomous way (Rosanvallon, 2009; Castoriadis, 1997). Are evaluation criteria democratically justified? Does the evaluation system embody a democratically appropriate balance between micro-quality issues and macro-quality issues? For example, with an over-focus on micro-quality, the evaluation system collects enormous amounts of information about implementation and management issues, whereas there is limited evaluation of policy decisions. Does the evaluation also embody a democratically justified balance between defensive quality and offensive quality, where defensive quality focuses on adherence to standards and avoidance of risks, and where offensive quality stands for risk-taking and innovation?

**19**. Does the evaluation system incorporate learning mechanisms and ways to ensure a responsiveness to critique that are meaningful and appropriate compared to the institutional power invested in evaluation systems? If ongoing learning and responsiveness are expected, these properties should be given reflected in the way the evaluation system is designed.

## Perspectives and conclusions

Further research may help make the proposed list of questions in EA 2.0 more systematic and rigorous, and better integrated into a theoretical framework.

Nevertheless, EA 2.0 is a preliminary way to talk about how evaluation systems can be more reflexively and thoughtfully implemented. It represents a healthy anti-dose to a belief that evaluation is a universal good. EA can be used in a constructive dialogue between politicians, administrators, consultants, citizens and others to gauge the usefulness of evaluation systems in particular situations.

However, no version of EA is easy. It cannot be reduced to a narrow algorithm limited to a few decisions in the early phases of building an evaluation system. EA 2.0 offers a broad and holistic perspective on the situational usefulness of an evaluation system which may be especially helpful in the early phases, but which should not be forgotten as the life of the evaluation system unfolds in practice. In contradistinction to original EA, EA 2.0 looks more at the evaluation (system) itself, and it continues to look at the actual function of that system also after its implementation. A processual perspective is called for, given the complex interaction between evaluation systems and their political and organizational contexts, and given their large-scale nature in the middle of a dynamic knowledge society. EA 2.0 is thus likely to produce less clear-cut results than did the original EA, and it is less easy to isolate as a distinct ex-ante procedure.

EA 2.0 is up against the interests of the evaluation industry. It is not comfortable if EA, old or new, suggests that evaluation is not appropriate for the time being. It is unbeneficial for evaluators to decline a commission, which they ought to do if the outcome of the EA is negative (Shadish, Cook & Leviton, 1991: 237). Wholey´s (2004) concern—that evaluation and EA should be as least costly as possible—is sympathetic, but not in the interests of the evaluation industry. Instead of facing a strictly rational set of entry criteria before selling one evaluation, consultants, inspectors and evaluators are now in position to sell a whole culture of evaluation to the extent that evaluation capacity/culture/systems are accepted as generally good ideas. And with the belief that evaluation systems should be institutionalized, a large number of well-paid jobs in evaluation centres and in the leading administrative and managerial circles will be secured, too. In the era of evaluation systems, perhaps the function of evaluation is so tightly integrated in political and administrative systems that an independent evaluability assessment is not relevant or not demanded by anyone. With the mainstreaming of evaluation into steering, public administration and management regimes, it becomes more unlikely that the evaluation field will define its own services as situational goods. They are more likely to be sold as universal goods, and constantly be extended across time and space in a "systematic" way, although this approach does not always include a careful estimation of how much evaluation systems cost (Power 1997) and what their effects are.

There is also politics, of course. Even a careful EA 2.0 may underestimate the extent to which decision makers may want to use an evaluation system to promote a particular agenda regardless of how little that some think that the system "fits" the situation at hand. Politics itself will seek to define a given social and political situation. If evaluation systems are integrated into a political agenda which emphasizes governance and control as purposes in themselves, then evaluation systems may serve their purposes even if they do not enhance reflexivity, learning, or improvement. To the extent that evaluation systems are seen as embodiments of New Public Management assumptions and beliefs, this discussion is important.

Nevertheless, exactly because of its rational overtones, EA 2.0 may be a promising idea in situations where there is a political, social and organizational preparedness at least to check or discuss whether the belief in evaluation has become an ideology or whether evaluation is likely to deliver what it promises – *under specific circumstances at hand*. The democratic perspective in EA 2.0 is not negligible. If the consequences of evaluation systems are spreading in society because evaluation systems are, it may be an important democratic task to discuss the circumstances under which evaluation systems are appropriate or not. EA 2.0 provides one way of framing an argument about this issue.

A careful EA will continue to struggle with the ideal that evaluation should be based on rational decisions and the knowledge that it will not be so in reality. It would be the mother of all paradoxes if EA in any version became a mandatory and comprehensive checklist that should be adhered to in all situations.

## References

Abma, Tineke & Mirko Nordegraaf (2003) 'Public Managers Amidst Ambiguity: Towards a Typology of Evaluation Practices in Public Management', *Evaluation* 9(3): 285-306.

Baizerman, Michael, Donald W. Compton & Stacey Hueftle Stockdill (2005) 'Capacity Building', in S. Mathison (ed.) *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage.

Boltanski, Luc & Eve Chiapello (2007) *The New Spirit of Capitalism*. London: Verso.

Castoriadis, Cornelius (1997) *World in Fragments. Writings on Politics, Society, Psychoanalysis, and the Imagination*. Stanford: Stanford University Press.

Dahler-Larsen, Peter. (2012). *The Evaluation Society.* Stanford: Stanford University Press.

Fitzpatrick, Jody L., James R. Sanders & Blaine R. Worthen (2004) *Program Evaluation: Alternative Approaches and Practical Guidelines.* Boston: Pearson Education, Inc.

Hood, Christopher (2002) 'The Risk Game and the Blame Game'. *Government and Opposition* 37(1): 15-37.

Leeuw, Frans L. & Jan-Eric Furubo (2008) 'Evaluations Systems: What Are They and Why Study Them?', *Evaluation* 14(2): 157-169.

Leviton, Laura C., Charles B. Collins, Beverly L. Laird &Polly P. Kratt (1998) Teaching Evaluation Using Evaluability Assessment', *Evaluation* 4(4): 389-409.

Meyer, John W., John Boli & George M. Thomas (1994) 'Ontology and Rationalization in the Western Cultural Account' in W.R. Scott and J.W. Meyer (eds.) *Institutional Environments and Organizations*, pp. 9-27. Thousand Oaks, CA: Sage.

Munro, Eileen (2004) 'The Impact of Audit on Social Work Practice', *British Journal of Social Work* 34: 1075-1095.

Pollitt, Christopher (1995) 'Justification by Works or by Faith? Evaluating the New public Management', *Evaluation* 1(2):133-154.

Pollitt, Christopher (2006) 'Performance Information for Democracy. The Missing link?' *Evaluation* 12(1): 38-55.

Pollitt, Christopher (2010) *Accountability: A concept that has expanded so much it may burst?* Paper to support keynote speech at Riksrevisionsdagen 2010: kunskap och accountability, Stockholm, 12 April 2010.

Power, Michael (1996) 'Making Things Auditable', *Accounting, Organizations and Society* 21(2/3): 289-315.

Power, Michael (1997) *The Audit Society*. Oxford: Oxford University Press.

Power, Michael (2005) 'The Theory Of The Audit Explosion' in E. Ferlie, L.E. Lynn and C. Pollitt (eds.) *The Oxford Handbook of Public Management,* pp. 327-344. New York: Oxford University Press.

Rist, Ray C. & Nicoletta Stame (2006) *From Studies to Streams*. New Brunswick: Transaction Publishers.

Rosanvallon, Pierre (2009) *Demokratin som Problem*. Hägersten: Tankekraft Förlag.

Rossi, Peter H., Howard E. Freeman & Mark W. Lipsey (2004) *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage.

Rothstein, Henry, Michael Huber &George Gaskell (2006) 'A Theory of Risk Colonization: The Spiralling Regulatory Logics of Societal and Institutional Risk', *Economy and Society* 35(1): 91-112.

Røvik, Kjell Arne (1998). *Moderne Organisasjoner. Trender i organisasjonstenkningen ved tusenårsskiftet*. Bergen-Sandviken: Fagbokforlaget.

Schwandt, Thomas A. (2009) 'Globalization Influences on the Western Evaluation Imaginary' in K.E. Ryan and J.B. Cousins (eds.) *The Sage International Handbook of Educational Evaluation,* pp. 19-36. Thousand Oaks, CA: Sage.

Scott, William Richard (1987) 'The Adolescence of Institutional Theory', *Administrative Science Quarterly* 32: 493-511.

Scott, William Richard (1995) *Institutions and Organizations*. Thousand Oaks, CA: Sage.

Shadish, William R., Thomas D. Cook & Laura C. Leviton (1991) *Foundations of Program Evaluation: Theories of Practice.* Newbury Park: Sage.

Smith, Midge F. (2005) 'Evaluability Assessment'. In S. Mathison (ed.) *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage.

Stehr, Nico (1994) *Knowledge Societies*. London: Sage.

Trevisan, Michael S. (2007) 'Evaluability Assessment From 1986 to 2006', *American Journal of Evaluation* 28(3): 290-303.

Trochim, William M.K. (2009) 'Evaluation Policy and Evaluation Practice', *New Directions for Evaluation* 123: 13 – 32.

van Thiel, Sandra &Frans L. Leeuw (2002) 'The Performance Paradox in the public Sector', *Public Performance and Management Review* 25(3): 267-281.

Vedung, Evert (2010) 'Four Waves of Evaluation Diffusion', *Evaluation* 16(3): 263-277.

Wholey, Joseph S. (2004) 'Evaluability Assessment' in Joseph S. Wholey, Harry P. Hatry & Kathryn E. Newcomer (eds.) *Handbook of Practical Program Evaluation*, pp. 33-62. San Francisco: Jossey-Bass.

Wholey, Joseph S. (1987) 'Evaluability Assessment: Developing Program Theory', *New Directions for Program Evaluation* 33 Spring: 77-92.

## Notes

[1] Whether Wholey´s idea of EA was based on a particular assumption about the form of evaluation to be carried out will be taken up in the next section. Instead, we focus on the most useful aspect of EA which is the idea that it is possible ex ante to determine whether evaluation is a good thing under articular circumstances.

[2] This outcome constitutes a critical moment like Popper´s famous falsification of a theory. If the idea that evaluation is good is not to become a dogma, there must be a least one thinkable set of circumstances where EA says "but not in this situation."

[3] His statement is substantiated by a search in Social Science Citation Index, which I did and can recommend to the reader. While the number of texts on evaluation is growing, as is the number of texts on audit, performance management, performance indicators, and accreditation, the number of texts on evaluability assessment is low and not growing. Critics will argue that SSCI is not the best place to check the popularity of EA. But why not, if evaluation, performance management, performance indicators etc. are discussed here?

[4] A holistic EA 2.0 of an evaluation system is useful even if the system has already been put in place based on a violation of one of the earlier requirements of EA. In addition, on Fitzpatrick et al.´s list, the first item is "Is there a legal requirement to evaluate?". If yes, it is recommended to skip the rest of the EA and to go directly to evaluation. But even a legally mandated evaluation may benefit from the thoughtfulness that flows from a more comprehensive EA.