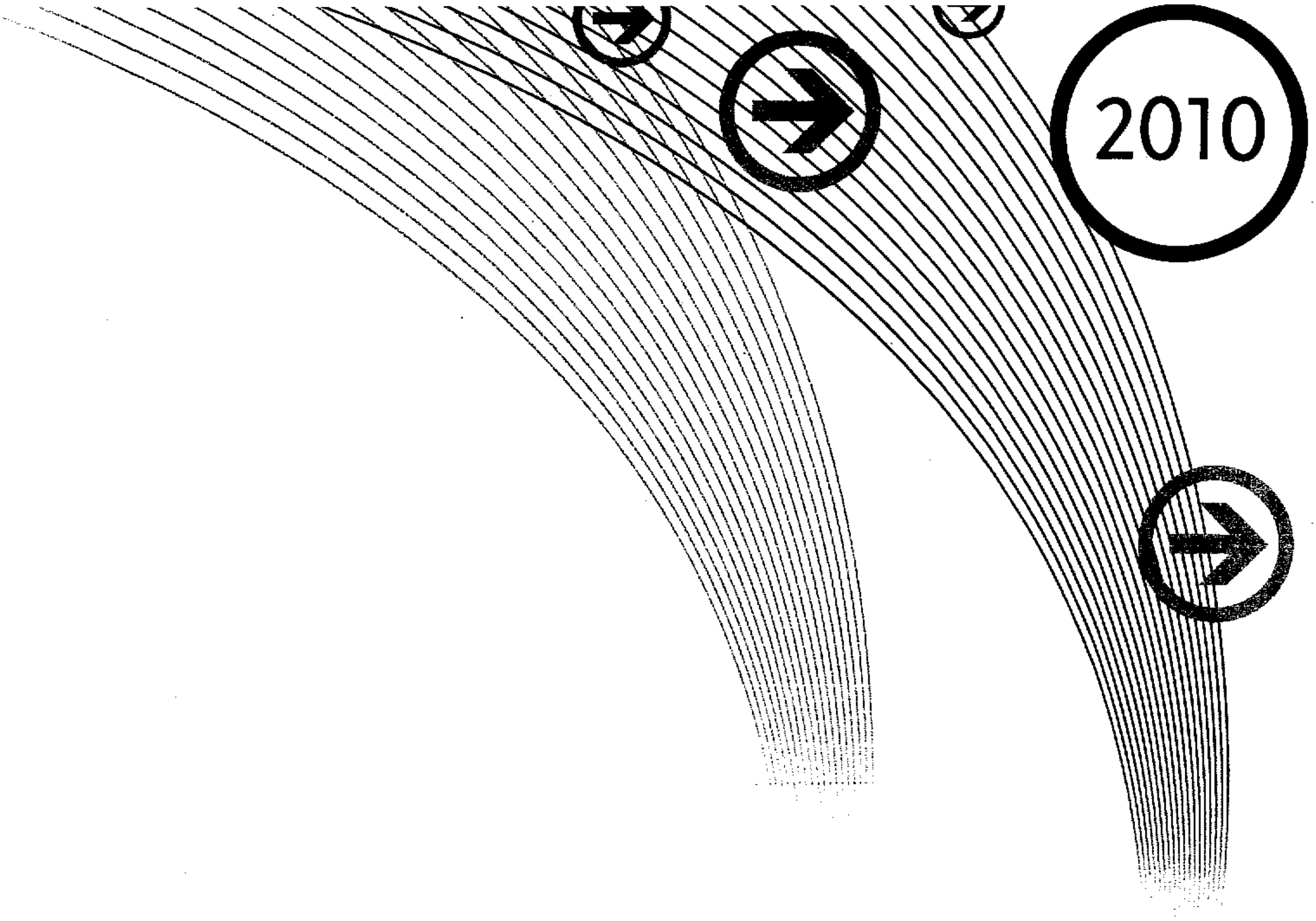

Riksrevisionsdagen

kunskap och accountability

RIKSREVISIONEN

> ➔ *"Alltför ofta ropas på utvärdering utan att hänsyn tas till den specifika situationen – att utvärdera har kommit att ses som en "universal good", något som rent normativt alltid är bra oavsett förutsättningarna."*

Peter Dahler-Larsen – Om att bedöma vad som ska utvärderas

VIKTEN AV ATT BEDÖMA VAD SOM SKA UTVÄRDERAS

Kort referat

Statsvetaren Peter Dahler-Larsen, professor vid Syddansk universitet, höll ett föredrag med titeln "Evaluability Assessment 2.0 - Or: On a Reflexive Approach to Evaluation Systems in the Knowledge Society". Dahler-Larsen menade att det alltför ofta ropas på utvärdering utan att hänsyn tas till den specifika situationen – att utvärdera har kommit att ses som en "universal good", något som rent normativt alltid är bra oavsett förutsättningarna. I polemik mot denna syn levererade Dahler-Larsen en lista på kriterier som bör tas in i bedömningen av vad som ska utvärderas, som sammantaget kan ses som en verktygslåda för *Evaluability Assessment*. Denna form av bedömningar har enligt Dahler-Larsen helt upphört sedan 70-talet, vilket är paradoxalt med tanke på att samhällets intresse för utvärderingar har vuxit dramatiskt under samma period.

I Dahler-Larsens "EA 2.0" ingick bland annat:

- *Vad* utvärderas? Vissa saker utvärderas ofta och kontinuerligt – u-hjälp eller socialt arbete – medan andra företeelser – krig eller kungafamiljer – aldrig tycks vara i behov av utvärdering.
- *Vilka mål* har det som utvärderas? Det går inte att utvärdera en verksamhet som saknar ett tydligt mål eller där det saknas konsensus om vad målet är.
- Är *strukturen* på det som ska utvärderas lämplig för en integrerad utvärdering? Om verksamheten är komplex kan det vara rimligare att utvärdera de enskilda delarna utifrån de infallsvinklar som är tillämpbara för varje enhet, snarare än att analysera helheten utifrån ett integrerat utvärderingssystem.
- Är verksamhetens *praktiska* betydelse för användarna av en sådan art att utvärderingen av enskildheter är relevant? Till exempel: läkare i Danmark hålls ansvariga för sina journaler. I en lång sjukdomsprocess finns ofta ett stort antal journalanteckningar. Är det i patientens intresse att varje enskild doktors anteckning utvärderas? Ibland kan utvärderingssystem överbetona ansvarsutkrävande på mikronivån (*micro-accountability*) på ett sätt som gör att för lite uppmärksamhet riktas mot det långsiktiga och övergripande målet med verksamheten.
- Har alternativ *kunskap* beaktats? I kunskapssamhället står sällan valet mellan utvärdering å ena sidan och total okunskap å den andra. Snarare finns det ofta redan flera alternativa och kompletterande kunskapskanaler. Kanske granskas redan verksamheten av ett annat organ?

Vad är existensberättigandet i att flera organisationer granskar samma sak? Har en utvärdering ett berättigande – eller bidrar den enbart till ytterligare överproduktion av underkoordinerad kunskap?

- Är *tekniken* som ska användas för utvärderingen tillräckligt pålitlig? Till exempel: breda datoriserade tester av skolelever, som görs i Danmark, förutsätter naturligtvis att den digitala infrastrukturen är funktionerlig.
- Vad är *kostnaden* för att utvärdera – både för den granskande parten och för den som granskas i form av tid, merarbete och förberedelser?
- Motiveras utvärderingen av *ideologiska* snarare än praktiska faktorer? När det är värderingar och normer som ligger till grund för utvärderingar, till exempel för kvalitetsutvärderingar i sjukvården, riskerar utvärderingen att bli ett självändamål snarare än ett praktiskt verktyg. Utvärderingen förlorar då snabbt legitimiteten hos dem som utför den.
- Finns det *alternativ* till att utvärdera? Går det att höja kvaliteten på andra sätt?

Under den följande frågestunden konstaterade professor Christopher Pollitt att han delade Peter Dahler-Larsens synpunkter, liksom hans skepsis mot de obligatoriska, övergripande utvärderingar som genomförs slentrianmässigt och utan selektivitet. Han betonade att denna inställning bygger på en vilja att försvara utvärderingsmekanismer, men genom att se till att de används korrekt och inte "rullas ut som en filt" vilket ger dem dåligt rykte och motverkar sitt syfte.

Evaluability Assessment 2.0. Or: On a Reflexive Approach to Evaluation Systems in the Knowledge Society

Peter Dahler-Larsen

Introduction

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

The field of evaluation has developed a heuristic tool called "evaluability assessment" (EA) which is supposed to help evaluators decide upon that question. EA incorporates a rationalistic view of knowledge according to which each piece of knowledge should be bought and rational decisions should be made about the following step.

In some interpretations of EA, the idea is that some programs 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 not be applied under all circumstances. EA is an expression of the latter of these beliefs.

The idea of EA was popular in the 70'ies, but the idea largely died out for a number of reasons. The broader values, norms and assumptions connected to the practice of evaluation in our society, ie. what Schwandt (2009) calls the "evaluation imaginary", now supports the idea that evaluation is a *universal good*. The perceived need to check whether it is appropriate under particular circumstances disappears. Perhaps that is why the promotion of evaluation culture, evaluation capacity, and evaluation systems (Leeuw and Furubo 2008) has gained ground in the broader social context of "The Audit Society" (Power 1997).

Could EA be re-vitalized, perhaps leading to a more modest, more reflexive, and more context-sensitive belief in evaluation? However, a re-invention of EA is only serious if it takes two contemporary phenomena into account: First, the complexities connected to the role of knowledge in the contemporary *knowledge society*, and second, the organizational and social properties of generic, repetitive, mandatory, and comprehensive *evaluation systems* (Leeuw and Furubo 2008) instead of just stand-alone evaluation studies.

The role of knowledge in contemporary society

More than anything else, three characteristics stand out as defining aspects of the knowledge society.

First, "the knowledge society" refers to an increased production of knowledge in society, including a broad diffusion of the capacity to produce knowledge (so that knowledge is no longer under the reign of select monopolies such as universities and state authorities). The producers of knowledge about the public sector include a wide multitude of knowledge centers, quality assurance institutions, evaluation centers, think tanks, universities, students, consulting companies, international policy-making organizations, standard-setting organizations, NGO´s, and Supreme Audit Institutions. Since there is a multitude of these actors, we can assume that their perspectives and their produced "knowledges" sometimes stand in a disorganized, competitive, and/or reflexive relation to each other.

Second, and perhaps more subtle, knowledge plays a new role in relation to the social order itself, ie. a more socially productive role (Stehr 1994: 103). As the social order becomes exposed to various forms of knowledge (organizational knowledge, market analysis, economic models, therapeutic inquiry, medical diagnoses, public inspection, quality control and audit etc) the social order is changing itself in the light of the knowledge produced. For example, knowledge defines categories (such as diagnoses) based on which social action takes place. Knowledge draws attention to areas of social controversy. Knowledge delivers critique of the existing conditions. Knowledge transports ideas, problems and proposed solutions across social contexts. Knowledge suggests how the social order should be changed. And new initiatives seek legitimacy with reference to knowledge. In other words, as society uses knowledge in its self-reflexivity and self-organization, the social order becomes fragile (Stehr 1994). More activities, structures, programs, organizations and initiatives become unstable because knowledge acts upon them.

Third, às Stehr (1994) immediately points out, however, knowledge as a capacity for action upon the social order too often simplistically assumes a bureaucratic or otherwise smoothened order with linear structures which are "prepared for data processing" (p. 103). This assumption rarely holds. A number of factors help explain why the transmission of knowledge into changing social relations is far from linear and straightforward. The first is the unequal and unstable distribution of power. Another is that changes in values follow other logics than changes in the systems which produce knowledge. Knowledge today must not only be (relatively) true, it must also be meaningful, appropriate, acceptable, legitimate and perceived as relevant (Gibbons et al. 1994). Still another factor is the vast influence from unintended consequences of earlier applications of knowledge. The reflexivity of modern social relations is itself a non-linear and thus de-stabilizing factor (Giddens 1994: 44-45). Therefore, knowledge produces opportunities for planning and direct regulation relatively rarely, whereas the typical result is *an increasingly fragile social order, but also one in which an element of unplanned events is ever-present.*

Against this view one might argue that if knowledge regimes can be kept relatively stable, perhaps based on power, then the resulting social order will be correspondingly more solid. Still, Stehr's and Giddens' point is a healthy antidote to an overly instrumental view on the construction and use of knowledge. Instead, the use of knowledge is a dynamic and sometimes surprising social and democratic learning process rather than as a predictable, linear and instrumental process.

Summing this section up, we have three observations:

- the capacity to produce knowledge has multiplied in society
- knowledge leads to a fragile social order
- the impact of knowledge unfolds in non-linear ways

Evaluation systems

There has been a paradigm shift in the social organization of evaluation. The classical paradigm for evaluation as an in-depth, expert-based, ad-hoc inquiry. Today, however, more emphasis is on evaluation systems (Leeuw and Furubo 2008). Evaluation systems are fairly permanent, repetitive, and routine-based. They 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, standards, benchmarks 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 and 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 and Furubo 2008).

To an increasing extent, evaluation systems are supported by power regulatory institutional pillars (Scott 1995) as that they do not only perform an information function, but also a resource allocation function (e.g. according to New Public Management prescriptions) and/or a legal function (as in some mandatory accreditation systems and many audit systems).

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 2005), to enhance evaluation cultures, and to create systematic managerial and organizational approaches to ongoing evaluation so that evaluation could be 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 and 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 and fairly vague assumptions about the virtues of particular organizational recipes (Røvik 1998; Meyer, Boli and Thomas 1994). By focussing on or installing evaluation systems, the burden of data collection may also be decreased, or streamlined, at least from the viewpoint of an inspector or evaluator.

This does not necessarily mean that evaluation systems are not required to produce enormous amounts of data. The point is that the burden placed on the inspector or evaluator is relieved. This is done by requiring the inspected organization to produce the necessary documentation and/or by raising the level of abstraction on which the inspector or evaluator operates. Whether "the quality system is in place" or not can be turned into an operational question in the eyes of the inspector. The abstraction from "things done" to "systems controlling how things are done" is also highly beneficial for the evaluator/inspector because his/her expertise is – at best – only general and abstract.

The institutional advantages (or potential disadvantages) of evaluation systems therefore 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 external auditability (Power 1996: 300).

On top of these concerns, a broader set of social norms seem to be consistent with the mentality inherent in evaluation systems.

In Power's (1996) analysis, the self-gratulatory process of evaluation checking evaluation 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 in the public and private", says Power (2005: 341).

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.

The evaluation industry as such is not without interest in this change, and it 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 evaluation, whether or not the prizes to be won is profit or institutional power.

The interest in EA has waned among evaluators, partly because it is unbeneficial to decline the commission, which they ought to do if the outcome of the EA is negative (Shadish, Cook and Leviton 1991: 237). Wholey´s classical 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 not only an evaluation but a whole culture of evaluation to the extent that evaluation capacity/culture/systems are accepted as generally good ideas.

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 doing so, various techniques/approaches/strategies who may not all have originated as "evaluation" have been able to "generalize" themselves as broadly applicable social practices (Power 1996). For example, quality inspection has moved from a fairly technical domain (dominated by engineers) to a more general managerial domain (Power 1996: 300), first in the industrial sector, then in the service sector, and lastly more generally in society. Now, "quality" is the overarching headline for public reforms in several countries. In a similar vein, audit has moved from a strictly financial domain into broader organizational systems, partly blending with quality inspection, quality assurance etc. (Power 2005: 333). And not surprisingly, accreditation is no longer a specific procedure used by insurance companies to check whether organizations can be insured, but rather a general formula for official, authoritative approval of an organization or institution. In other words, the "field of evaluation" is not a socio-historical constant. It has managed to grow and expand through the adoption and integration of a number of data-producing practices which have generalized themselves at the same time as they shifted their focus from looking at things to inventing systems which look at systems which do things.

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 and 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 and Leeuw 2002), and that evaluation systems have constitutive effects on practice. Constitutive effects refer to the ways in which evaluation systems shape behavior and redefine the meaning of public activities because evaluation indicators become goals in themselves (Dahler-Larsen 2007).

The finding that evaluation systems have such effects is very consistent with our observations about the contemporary role of knowledge. Knowledge intervenes in the social order and cannot be kept separate from it; but the impact of knowledge-producing systems on that order is non-linear and often characterized by paradoxes and unintended effects.[1]

If there is more than a grain of truth in this argument, it may be fruitful to encourage further reflexivity among the architects of evaluation systems. It may be fruitful not only to question the design of such systems, but also to distinguish better between situations where such systems are very appropriate and situations where they are not. A continuous reflexive view upon the role and function of evaluation systems may also be fruitful it is true that their effects are not fully predictable. In other words, it may be fruitful to re-vitalize and update the idea of EA. Our update will be called EA 2.0. But let us first recapture EA 1.0.

Evaluability Assessment 1.0

EA is a process which leads to a decision about whether it is sensible to evaluate under given circumstances. The idea is practical and useful if one wants prescriptions about when to evaluate and when not to. But *the idea is not nearly as popular today as it was in the 70´ies* (Smith 2005: 137). 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, EA may lead to a negative conclusion. But let´s unfold the argument from the beginning.

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 and Leviton 1991: 237).

According to the idea of EA, a number of circumstances in an evaluand and its context should be clarified before evaluation is undertaken. Potential problems with the program that makes it non-evaluable should be straightened out (much in the same way that a hair-dresser combs your hair before cutting it) before evaluation can proceed. As a consequence, it is possible that under some circumstances – when the "problems" of the program have not been straightened out or perhaps cannot be straightened out in the near future, or perhaps will never be straightened out – then evaluation may not be the best idea in the world. Resources may then be spent better on evaluating other things or on other things than evaluation.

Now, however, let us look at the strict logic of EA. The following questions should be answered (Shadish, Cook and Leviton 1991: 237; Rossi, Freeman and 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. Is the program fairly well implemented? If it is known that it is not, it is wiser to improve the implementation of the program before evaluating it. 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. And 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).
c. Is there a fairly good program theory? If not, it is better to clarify the logic of the program and perhaps improve it accordingly before evaluation is undertaken.

[1] This analysis of evaluation systems is also in line with the main conclusions of institutional theory in organizational analysis: Perhaps these systems are adopted not because they are superior instruments to the handling of technical problems, but because they reflect broader prevailing norms, values, and myths in society.

d. Are there well-described and plausible goals? If not, the outcome of the evaluation is 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 not in position to make a practical difference.

In essence, the idea of testing a program's evaluability is highly rational. It assumes that evaluation can be deliberately applied strictly to those situations in which it can make an instrumental difference. Evaluation should be carried out only when appropriate, and there are fairly strict operational definitions of this appropriateness.

Attempts have been made to invigorate EA since its golden age in the 1970 'ies and it is still promoted by some, but there is no abundant and booming literature about the topic. A literature search demonstrates that while the literature on evaluation has grown since the 1990ies, the literature on EA is limited and has not grown (see appendix 1), although it may survive in some disciplines rather than in evaluation as such (Trevisan 2007). Why not in evaluation, if it is a good idea?

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, it would be unrealistic to expect that a universal algorithm for EA could be developed, and the spirit of EA should thrive even without a strict algorithm.

Next, EA as defined above may be appropriate only if the intended subsequent evaluation is an old-school goal-based evaluation assuming clear goals and proper means-ends relations. Since the birth of EA in the 1970 'ies, 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 conflicting and can be dealt with during the evaluation itself.

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) separate from evaluation as such.

Perhaps the most deeply lying assumption in the evaluation imaginary connected to EA is 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 its underlying assumptions of strict rationality, EA has declined since the 70 'ies (Smith 2005: 139). If our analysis in the previous section about the social roots of evaluation systems is correct, there may be normative or ideological reasons for evaluation that are not always rationally justified.

The ideology of evaluation may have moved from regarding evaluation as an instrumental situational good to a universal good. In an era of evaluation culture, - capacity and -systems, a singular evaluation does no longer have to demonstrate its rational benefits. It is the belief in the long-term principle of evaluation that counts. The elimination of EA makes sense if evaluation culture, capacity and systems are recommended regardless of an analysis of each specific evaluation. It is not in the best interests of the evaluation industry to run into situations where EA suggests that evaluation is not appropriate for the time being. It is better for the industry to promote the idea that evaluation is a general good. With the development of evaluation systems in all their varieties, there is a broad and expanding market for evaluators interested in various aspects of evaluation, such as audit, organizational learning, management information systems, organizational development, statistics, reporting and communication etc. And with the belief that evaluation systems should be institutionalized, a large number of well-paid jobs in the leading administrative and managerial circles will be secured, too.

With EA being (largely) out of the picture, the road has been paved for an ever-increasing social self-confidence of evaluation. Evaluation does not need to argue for itself in each and every situation. It may be sound, however, to re-introduce a healthy dose of self-reflection into the contemporary belief in comfort-producing evaluation systems. The following is in line with that spirit - by suggesting a re-formulated and updated EA 2.0 for the era of evaluation systems.

Evaluability Assessment 2.0

A few of the items in EA 2.0 are repetitions or slight revisions of the original questions, now adapted to evaluation systems. Some of the elements in EA 2.0 may be fairly pedestrian pieces of advice, or repetitions of good advice offered in other works (Fitzpatrick et al. 2004; Wholey 2004), but there are also a few points which are unique to the perspectives on knowledge offered above.

The items in the following EA 2.0 are merely organized as a list of factors which deserve to be considered. I have much sympathy for a logical, stepwise algorithm, where a "no" to item one cancels all further down the list (see eg. Fitzpatrick et al. 2004), but I have not copied such approach, because a holistic all-things-considered-approach has some critical advantages.[2]

Characteristics of the evaluand

1. Does the object of evaluation have enough social impact or importance to warrant a formal evaluation system? (Fitzpatrick et al. 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, and public management ideologies. For example, New Public Management initiatives have not been evaluated nearly as much as they deserve (Pollitt 1995).

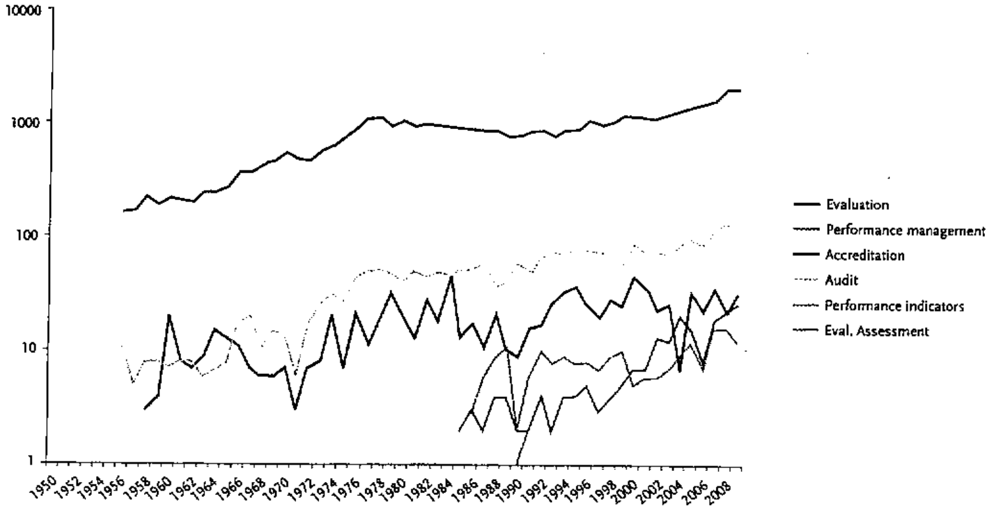

Figure 1: Social Science Citation Index (SSCI) Search of March 24, 2010. Title=("begreb") AND Topic=("begreb") Refined by: Languages=(ENGLISH) Timespan=1955-2009. Databases=SSCI.

[2] 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 comprehensive EA.

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 and Nordegraaaf 2003). The latter dimension refers to whether the user plays a substantial role in the successful outcome of the public service. Two-sided activities are for example therapy, learning (not teaching!), and prevention of risky health behaviour.

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 can be said to 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 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). In some accountability structures, 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, and the social responsibility for their solution cannot 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 overemphasize 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 and 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 centers, accreditation centers, 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 together overproduce knowledge that is under-coordinated.

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. Each year, teachers and pupils have prepared for the national computerized testing, and subsequently Danish citizens can read in the news about stalled computers, black screens, cancelled tests and frustrated pupils, teachers, and principals. The company delivering the digital infrastructure has simply not succeeded in developing a system that can work reliably in the national scale. EA would suggest that the evaluation should be made functional and then national, not national and then functional.

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 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 knowledge generally is in the knowledge society.

10. Is the evaluation system able to provide dependable information? (Fitzpatrick et al. 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. Costs. 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 documentation 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.

An account of costs is necessary if a calculation is to be made which 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 times 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.

12. Ideology and self-representation. Is the evaluation system infused with an overarching ideology that tends to make the evaluation system self-justifying? Does the ideology of that evaluation system match the actual characteristics of the evaluation system and the context in which it operates? 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. In a culture favourable to evaluation, consensus may reproduce social myths about what is "best practice" and how the fancy new evaluation system leads to "quality" and "learning" and "improvement" and "better services."

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 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 of the non-linearity issue in the use of knowledge.

If there is not agreement among key stakeholders about the intended use of the revaluation system (Fitzpatrick et al 2004: 186) 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).

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 claim 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 no convincing 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 likely to be 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 in practice.

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 repetitive, 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 may have benefitted from pilot testing.

15. Have the consequences of the evaluation system (apart from its official purpose) been investigated? In an EA perspective, it is beneficial to ask: How are people under the evaluation system likely to behave if they take the evaluation criteria seriously but have little supplementary guidance? 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.

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 and 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. In the knowledge society, it should be regarded 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 issues 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. Learning mechanisms and responsiveness to critique. 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?

Perspectives and conclusions

EA is a way to talk about how evaluation systems can be more reflexively and thoughtfully used in a societal situation where knowledge plays an increasing role in ways that are often not planned and intended. EA is also a healthy anti-dose to a belief that evaluation is a universal good.

However, EA even in version 2.0, is not easy. It cannot be reduced to a narrow algorithm limited to a few decisions in the early phases of building an evaluation system. Instead, EA 2.0 is a set of interrelated observation points that question the interaction between the evaluated activities, the evaluation system, and the broader social context in which the knowledge produced by the evaluation system exerts an influence. EA 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 a similar vein, EA cannot be the narrow responsibility of only a particular architect of an evaluation system. For EA to be meaningful in a complex knowledge society, it needs to be connected to a broader social process.

Still, a careful EA will continue to struggle between the ideal that evaluation should be based on rational decisions and the knowledge that it will not be so in reality. And 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 well that system "fits" the situation at hand. Opposition, struggle and controversy are the order of the public world in a way that is neither represented in EA 1.0 nor 2.0.

Nevertheless, exactly because of its rational overtones, EA 2.0, it may be a promising idea in situations where evaluation systems have become self-justifying, or in situations where there is a social and organizational preparedness at least to check whether the belief in evaluation has become an ideology or whether evaluation is likely to deliver what it promises – under specific circumstances. Yet, 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, T. and M. Nordegraaf (2003). Public Mangers Amidst Ambiguity: Towards a Typology of Evaluation Practices in Public Management. *Evaluation*, 9(3): 285-330.

Baizerman, M.; D.W. Compton, and S.H. Stockdill (2005). Article on Capacity Building. In S. Mathison (ed.), *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage.

Boltanski, L. and E. Chiapello (2007). *The New Spirit of Capitalism*. London: Verso.

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

Dahler-Larsen, P. (2007). Constructive Effects of Performance Indicator Systems. In S. Kushner (ed.), *Dilemmas of engagement: Evaluation Development under New Public Management and the New Politics*. New York: Elsevier.

Fitzpatrick, J.L.; J.R. Sanders and Blaine R. Worthen (2004). Program Evaluation: Alternative Approaches and Practical Guidelines. 3rd edition. Bonston: Pearson Education, Inc.

Furubo, J.E. (2006). Why Evaluations Sometimes Can't be Used - and Why They Shouldn't. In R.C. Rist and N. Stame (eds.), *From Studies to Streams* (pp. 147-168). New Brunswick: Transaction Publishers.

Gibbons, M. et al. (1994). *The New Production of Knowledge. The Dynamics of Science and Research in Contemporary Societies*. London: Sage.

Giddens, A. (1994). *Modernitetens Konsekvenser*. København: Hans Reitzels Forlag.

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

Leeuw, F.L. and J.-E. Furubo (2008). Evaluations Systems: What Are They and Why Study Them? *Evaluation*, 14(2): 157-169.

Meyer, J.W.; J. Boli, and G.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, E. (2004). The Impact of Audit on Social Work Practice, *British Journal of Social Work* 34, 1075-95.

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

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

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

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

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

Power, M. (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, R.C. and N. Stame (2006). *From Studies to Streams*. New Brunswick: Transaction Publishers.

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

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

Rossi, P.H.; H.E. Freeman, and M.W. Lipsey (2004). *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage.

Røvik, K.A. (1998). *Moderne Organisasjoner. Trender i organisasjonstenkingen ved tusenårsskiftet*. Bergen-Sandviken: Fagbokforlaget.

Schwandt, T.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, W.R. (1987). The Adolescence of Institutional Theory. *Administrative Science Quarterly*, 32: 493-511.

Scott, W.R. (1995). Institutions and Organizations. Thousand Oaks: Sage.

Shadish, W.R.; T.D. Cook, and L.C. Leviton (1991). *Foundations of Program Evaluation: Theories of Practice*. Newbury Park: Sage.

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

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

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

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

Wholey, J.S. (2004). Evaluability Assessment. In J.S. Wholey; H.P. Hatry, and K.E. Newcomer (eds.), Handbook of Practical Program Evaluation (pp. 33-62). San Francisco: Jossey-Bass.

> *"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."*