The Power and Paradoxes of Evaluation Systems

Increasing Use but Impeding Change

Andersen, Niklas Andreas

Published in:
Scandinavian Journal of Public Administration

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

SJPA
25(3/4)

The Power and Paradoxes of Evaluation Systems – Increasing Use but Impeding Change

Niklas A. Andersen*

Niklas A. Andersen,
Department of Politics and Society, Aalborg University, Denmark
nia@dps.aau.dk

Abstract

In recent years, evaluation systems have become increasingly embedded within public sector organisations. This trend of systematising and institutionalising evaluation activities has generally been perceived as a way to increase the use – and thus the power - of evaluations. However, this article argues that the power of evaluation systems is of a more complicated nature than merely increasing the uptake of evaluative knowledge. By applying the concept of "contestability differential" to a concrete example of an evaluation system within the Danish employment services, it is shown how the institutionalisation of an asymmetric power relation between evaluation system and evaluand creates inherent paradoxes.

The analysis shows how the strong contestability differential between evaluation system and evaluand – necessary for securing the influence of evaluation systems - hinges on the permanence, organisational embeddedness and epistemological fixation of such systems. However, these same elements simultaneously also limit the usefulness of the evaluative knowledge and the capability of the evaluation system to invoke radical change and development in the evaluand.

The article thus highlights an inherent paradox of evaluation systems in that they are simultaneously increasing and decreasing the power of evaluations.

Keywords:
evaluation systems,
power,
paradoxes,
evidence-based policy and practice,
employment services

Introduction

One of the most significant current trends in the field of evaluation is the gradual institutionalisation of evaluation activities within the governments and public sectors of many different countries (Jacob, Speer and Furubo 2015; Stockmann, Meyer and Taube 2020). Evaluations are no longer primarily conducted as one-off studies, but are increasingly inscribed into larger streams of systematic knowledge production within specific organisations or organisational fields (Rist and Stame 2006). Concepts such as Evaluation Systems (Leeuw and Furubo 2008), Evaluation Machines (Dahler-Larsen 2012) and Evaluation Capacity (Bourgeois and Cousins 2013) have all been used to describe this shift in the nature and organisational embeddedness of evaluations. It is a shift that can be viewed as a response to the problem of non- or misuse of evaluations, which has haunted the evaluation field since its inception (Alkin and King 2016). The growing institutionalisation of evaluation activities into organisational procedures has thus hitherto mainly been studied as a means to increase the instrumental use and uptake of evaluations (Alkin and King 2018; Oliver et al. 2014). This is particularly evident in the comparative literature, where different countries are compared and/or ranked according to their degree of institutionalisation of evaluation activities within government and society (Jacob, Speer and Furubo 2015; Lázaro 2015; Rosenstein 2015; Stockmann, Meyer and Taube 2020), with the logic that the higher the degree of institutionalisation, the more evaluations are also used by governments and

Scandinavian Journal of Public Administration
25(3/4): 39 – 59

ISSN: 2001-7405
e-ISSN: 2001-7413

***Niklas Andreas Andersen** (PhD) is a Postdoctoral researcher at the Department of Politics and Society, Aalborg University. His area of research is situated at the intersection between Social- and Employment Policies, where he studies how techniques for evaluating, calculating and steering Public Policies are contributing to policy-change. He furthermore focuses on the institutionalization of Evidence-based Knowledge in the process of designing and implementing new welfare policies. He holds a cand.scient.adm. in Politics and Administration and a PhD in political science – both from Aalborg University.

more generally within society.

While these and similar studies provide valuable insights into ongoing processes of institutionalisation and systematisation of evaluation activities throughout the world, they are also somewhat limited by their view of evaluation activities as intrinsically beneficial. Adhering to this normative position skews attention from the many different and potentially adverse consequences of these institutionalisation processes towards a focus on how such processes can be furthered. The main exception is found within the burgeoning literature on systems of performance measurements, audits and quality assessments, which have dealt extensively with the more far-reaching and adverse consequences of such systems (for recent examples from the Scandinavian context see: Bjørnholt and Larsen 2014; Dahler-Larsen 2019a; Hanberger and Lindgren 2019; Segerholm 2020). These and many similar studies have amply demonstrated how the permanent and routinized production and dissemination of performance information can significantly alter how the subjects monitored are perceived and perform their tasks. However, within the evaluation literature such studies of performance monitoring are often regarded as being divorced from the practice of "true" evaluations; that is, policy and program evaluations (Furubo 2019). An illustrative example of this can be found in a recent survey of the institutionalisation of evaluations in Europe. The authors directly state that if the concept of evaluation is confused with audits, quality assurance etc., "(…) the fact-generating element of evaluation - so important for its use - is becoming lost and evaluation is getting entangled with institutional and managerial practices" (Stockmann, Meyer and Taube 2020, 500). In other words, any adverse or constitutive consequences of performance monitoring are unrelated to the practice of evaluation and, furthermore, such consequences are attributed to the entanglement with managerial practices rather than performance monitoring itself. The logic being that the institutionalisation and systematisation of program evaluations are not expected to cause the same consequences – especially not if evaluation researchers study how best to institutionalise and implement systems to increase the uptake of "the facts" generated by evaluations.

Contrary to this, this article argues that such consequences are also acutely present in the institutionalisation of evaluation systems and, furthermore, are related to the inherent nature of evaluations rather than (solely) being caused by factors external to the evaluation practice. The central claim is that the asymmetric power relation between evaluation and the subject/object of the evaluation (the evaluand) – so important in enabling evaluations to influence the evaluand – creates paradoxical situations when such power relations are institutionalised through evaluation systems, i.e. situations that ultimately make evaluation systems both increase and decrease the influence of evaluations. This paradox thus complicates the prevalent understanding of evaluation systems – and other forms of institutionalised evaluation activities - as a panacea to cure the ailment of non- or misuse of evaluations.

In substantiating this argument, the article draws on Peter Dahler-Larsen's concept of the "contestability differential" (Dahler-Larsen 2015a, 2015b), applying it to a concrete example of an evaluation system within the Danish employment services. The article thus seeks to both: 1) Expand our understanding of the nature and consequences of evaluation systems by

elucidating the power relations institutionalised through such systems, and 2) further theorise the contestability differential – and its analytical applicability to the study of evaluation systems - by introducing the notion of paradoxes.

The article proceeds as follows: The next section elaborates on the concepts of contestability differential and evaluation systems, arguing that the contestability differential has been used imprecisely in relation to evaluation systems and that this can be addressed by explicating the power dimension in the concept and introducing the notion of the paradox. This is followed by presentation of an evidence-based evaluation system in the Danish employment services, which functions as a heuristic to elucidate the article's theoretical arguments. This is achieved through a three-part analysis, with each part examining a different paradox related to the contestability differential – and the power relation inherent to this - created by the evaluation system. The final section summarises the article's main arguments and discusses the implications for the study of the institutionalisation and systematisation of evaluations.

Theoretical Framework

Contestability Differential

The "contestability differential" is a concept developed by Peter Dahler-Larsen to analyse the relationship between evaluation and evaluand (Dahler-Larsen 2015a, 2015b, 2019b). The concept is based on two important premises: (1) That evaluations always seek to change the evaluand; and (2) that both evaluations and evaluands are social constructions. The first premise should come as no surprise to most evaluation researchers, as this follows logically from most common definitions of policy and program evaluations. Take for example Evert Vedung's well-known definition, which states that evaluations are: "intended to play a role in future, practical action situations" (Vedung 1997, 3). If an evaluation did not seek to inform and thus possible alter the future state of the evaluand, it would not – by definition - be an evaluation.

The second premise is more controversial – at least viewed from the perspective of mainstream evaluation research. However, to say that something is a social construction can simply imply that it is actively formed out of disparate elements to make out a structure that is greater than the sum of its parts – much like the process of making a building (Hacking 2001, 50). But unlike a building, the objects which are social constructions are also held together by immaterial things such as ideas, language, intersubjective experiences etc. Following this line of reasoning, it should be clear that evaluations and evaluands are social constructions actively created and upheld through language (for example the definition of something as an "intervention" or a "performance indicator"), interaction (for example through the assignment of specific roles such as commissioner, evaluator, evaluand and user) and social imaginary (for example through ideals of accountability and effectiveness) (Dahler-Larsen 2015b).

These two premises have important analytical implications, as evaluations are understood as social constructions that deal specifically with other social constructions (the evaluands) in order to change these. Evaluations invoke this change in the evaluand by questioning its "merit, worth and value" (Scriven

1991,1) – i.e. by contesting the evaluand. However, to successfully question and contest the evaluand, the evaluation needs to be less contested than the evaluand – i.e. there needs to be a contestability differential. If the evaluation is more contested and less taken-for-granted than the evaluand, there would be no reason to give authority to the evaluation's claims about the evaluand. Dahler-Larsen uses the analogy of the relationship between a screw (the evaluand) and the person using a screwdriver (the evaluation) to explain the need for this difference in contestability:

> *"Assume someone is using force to turn a screw with a screwdriver. Imagine that the screw is solidly anchored and the connection with the screwdriver is strong, and the person has no solid position on the ground, then the force exerted will in fact lead to a turning of the person in space instead of a turning of the screw" (Dahler-Larsen 2015a, 31).*

Put in logical terms, the contestability differential of a given evaluation-evaluand relationship equals the contestability of the evaluand minus the contestability of the evaluation. Contestability is, of course, not a fixed and material thing that can be measured and given a precise numerical value, but the basic idea nonetheless remains the same; that a positive contestability differential is necessary for evaluations to be taken seriously and invoke change.

Such contestability differential is constructed through a process of solidifying the evaluation and/or destabilising the evaluand. The solidifying of the evaluation can be achieved through a number of means - for example by using sophisticated and not easily decipherable methods or by making evaluation a mandatory activity of organisations. The evaluand – whether being an actor (e.g. the performance of a person or an organisation), an object (e.g. a specific intervention or a policy) or a process (e.g. processes of decision-making or implementation) - can be destabilised and contested by for example being criticised for lacking transparency, failing to meet standards or not adhering to existing research-based knowledge (Dahler-Larsen 2019b).

Power, Contestability and Evaluation Systems

What then, can we learn by studying evaluation systems through the lens offered by the concept of contestability differential? In a seminal article by Frans L. Leeuw and Jan-Eric Furubo (Leeuw and Furubo 2008), the authors argue that evaluation activities form a coherent evaluation system when they meet the following four criteria: (a) evaluations are produced on a permanent basis and with a substantial volume; (b) the commissioning and/or production of evaluations are embedded within one or more of the central organisations in the given organisational field - rather than being the primary responsibility of external evaluators; (c) the evaluation activities are guided by a shared epistemological perspective; and (d) the system is geared towards securing continual availability of evaluative knowledge to be used in decision-making and implementation.

To make evaluation activities more systematic and institutionalised is thus to make the commissioning, production and use of evaluations part of an organisation's standard operating procedures – rather than an ad-hoc choice.

The four defining features of evaluation systems very much resemble the elements through which a strong contestability differential is often created – e.g. through mandatory processes, distinct methods and the backing of powerful organisations (Dahler-Larsen 2019b). This is perhaps unsurprising as the raison d'être of evaluation systems are to enhance the use and influence of evaluations, which directly hinges on creating a strong contestability differential. However, the relation between the concepts of contestability differential and evaluation systems – and the implications of this relation for the power of evaluation systems - have hitherto not been addressed within existing research.

Dahler-Larsen is, of course, well-aware of the potential power of evaluations, but he finds it problematic to - a priori - define evaluations as tools used by the powerful on the powerless, as he argues that evaluations can be used to serve the interests of any given actor and are also always at the risk of being challenged (Dahler-Larsen 2015a). He therefore highlights the concept of contestability differential as providing a tool for analysing the consequences of evaluations, without harbouring any preconceived notions of the power of evaluations.

While it is true, that the concept provides a much-needed non-normative perspective to analyse the consequences of evaluations, the above critique also leads Dahler-Larsen to downplay the power relation at the core of the contestability differential. The claim here is that the contestability differential is inherently a power relation between evaluation and evaluand. This is especially clear, if we apply the notion of power put forth by the French philosopher Michel Foucault. Foucault argued that power should not be understood as a fixed capacity, but should rather be understood in relational terms as a productive and reciprocal process of subjectification between the governing and governed subjects (Foucault 1990). Power is a process of subjectification in the dual sense of the word, as it both subjects people to different forms of control, but also makes people acknowledge themselves as specific subjects (Foucault 1982). Power is therefore productive as it not only limits, but also enables the actions of the governed. Power is furthermore reciprocal, as it is both the governing and the governed who are constructed as specific subjects through the relations of power.

Viewed through this prism, the contestability differential can be understood as a description of the power relation between a governing subject (the actors responsible for the commissioning, construction and/or production of the evaluation) and a governed subject (the actors responsible for the content, structure and results of the evaluand), which – through this relation – constitutes and structures the actions of both. At its core, the contestability differential is thus a power relation that can be more or less asymmetrical and skewed to either side of the evaluation-evaluand divide. This does in no way imply that evaluations are always powerful agents that shape evaluands, as this depends on the strength of the constructed contestability differential. However, it meana that for evaluations to achieve their defined purpose of shaping the future state of the evaluand, the power relation must successfully become skewed in their favour – i.e. through the construction of a strong contestability differential.

The point of the arguments above is not merely to explicate a power dimension already implicit within the concept of contestability differential. More

importantly, explicit use of the power terminology elucidates a number of – hitherto underexplored - paradoxes regarding the contestability differential of evaluation systems. A paradox can be defined as the self-contradictory presence of both parts of a distinction simultaneously – i.e. being a man and woman, a child and an adult, useful and useless etc. (Luhmann 1993). It is exactly such self-contradictory situations that the contestability differential of evaluation systems create by establishing certain power relations between system and evaluand.

As stated by Dahler-Larsen, the contestability differential is "an ever-present ingredient in evaluation" (Dahler-Larsen 2015a, 34), which – following the above arguments - means that an asymmetrical power relation between evaluation and evaluand is an ever-present ingredient of evaluation activities. Without it, we would not be able to evaluate our public policies and programs – i.e. reflect systematically on our agreed upon principles as a society. However, it is - in the words of Dahler-Larsen - important to make sure that any given contestability differential is only "preliminary, temporary and fragile" (Dahler-Larsen 2015a). Because to always let evaluations contest public policies and programs would be to permanently render policymakers and public servants powerless and thus unable to solve the problems at hand.

However, the idea of a fragile and temporary contestability differential runs into paradoxes when applied to evaluation systems, which seek to create contestability differentials that are fixed, permanent and strong. To make the contestability differential of an evaluation system "preliminary, temporary and fragile" is then – by its very definition - to hinder the system's functioning. On the other hand, to successfully establish a functioning evaluation system is to insert a permanent asymmetrical power relation between the actors responsible for the evaluations (whether these are commissioner, evaluator etc.) and the actors responsible for the evaluand (e.g. public organisations, frontline professionals etc.). By emphasising these power asymmetries, the current study of the contestability differential of evaluation systems thus eschews focus from the how and why of establishing the contestability differential, to questions regarding the consequences and paradoxes of institutionalising an asymmetrical power relation between evaluation and evaluand.

Case and Empirical Data

In order to flesh out the theoretical arguments above, the article draws empirical illustrations from the case of an evidence-based evaluation system within the Danish employment services. Similar to other countries and sectors, the idea of evidence-based policy and practice – i.e. developing policies and interventions on the basis of knowledge of 'what works' (Nutley, Walter and Davies 2003; Nutley et al. 2019) – gained traction within the Danish Ministry of Employment during the 2000s. Since the middle of the 2000s, this idea has gradually been institutionalised into the Ministry's internal procedures – culminating with the adoption of an official "evidence strategy" within the Ministry in 2012. Since then, the evaluative activities of the Ministry are structured within a coherent system, which resembles Leeuw's and Furubo's aforementioned definition of evaluation systems (Andersen 2020). This system is upheld through a

combination of: (1) an official knowledge hierarchy sanctioned by the Ministry of Employment, which places experimental and quasi-experimental impact-evaluations at the top; (2) routines within the Ministry for the continual commissioning, production and collection of such impact-evaluations; and (3) the construction of mandatory processes within the Ministry for assessing, accumulating and disseminating this knowledge to the relevant decision-makers and implementation agents (Andersen 2020). The analysis focusses on the relationship between the Ministry of Employment and the municipal job centres, as the former is responsible for the evaluations conducted by the system, while the latter is responsible for implementing the employment policies and programs being evaluated – i.e. the evaluand.

The empirical material has been collected through two research projects conducted between 2016 and 2020 (Andersen 2020; Andersen, Caswell & Larsen 2017; Andersen & Randrup 2017), both of which combined interviews (with managers, civil servants and caseworkers) and official documents (such as impact-evaluations, policy-documents and ministerial websites) from both the Ministry and seven different job centres. This large archive of qualitative data has been analysed through the prism of the concept of contestability differential, in order to derive illustrative and conceptually potent empirical examples of the power relation inserted by the evaluation system. These different examples were then compared and grouped into three overarching categories – each of which highlighted a different paradox created by the contestability differential instated by the evaluation system (cf. the following section).

The case is thus used instrumentally (Stake 2003) to provide novel insights into the concept of contestability differential and the inherent paradoxes of evaluation systems. To do so, the case has been selected on account of its uniqueness and richness of theoretical relevant information, rather than on it being a typical case of an evaluation system (Flyvbjerg 2006). The case is unique and especially informative because it encompasses all the defining elements of evaluation systems (Leeuw and Furubo 2008) to such a degree that it can be described as a highly institutionalised evaluation system (Andersen 2020). In other words, the evidence-based evaluation system is based on a very strong contestability differential, why the case should vividly illustrate the paradoxes of this situation – if indeed there are any. Furthermore, the case is also theoretical relevant in the sense that the evidence-based evaluation system is grounded in policy and program evaluations. The case can thus substantiate the article's initial claim about how the constitutive consequences of evaluations are not solely (or even primarily) caused by systems of performance monitoring (cf. section 1).

The Evidence-Based Evaluation System and the Paradoxes of Power

The following analysis will explore the contestability differential – and the asymmetrical power relation related to this - created by evaluation systems. This is done by zooming in on three different paradoxes, which are theoretically derived - by applying the lens of the contestability differential to the concept of evaluation systems – but empirically illustrated through examples from the

Danish case. Each of the paradoxes arises from one of the defining features of evaluation systems: 1) The permanence of evaluation systems; 2) The organisational embeddedness of evaluation systems; and 3) The epistemological coherence of evaluation systems.

The Paradox of Permanence

> *The permanence of evaluation activities makes the evaluand changeable, but limits the evaluand's ability to change*

The paradoxical statement above is grounded in the fact that evaluation systems make evaluation activities permanent in order to increase the availability and uptake of evaluative information in practical action situations. Evaluation systems thus render evaluands changeable by placing them in a permanently contested state, but such permanently contested state also makes it harder for the evaluands to change. Herein lies a paradox. To understand this paradox, it is necessary to understand the state of liminality caused by evaluations.

The concept of liminality was originally coined by the French ethnographer Arnold Van Gennep in his description of the rituals surrounding the rites of passages in tribal societies (Van Gennep 2019). Such rituals are, according to Van Gennep, typically structured into three phases: 1) The pre-liminal phase, where participants are separated from their normal environment and role; 2) the liminal phase, where the participants are placed in an indeterminate space between former and future roles; and 3) a post-liminal phase, where participants are reintroduced into their former environment, but now assuming a different role. As argued by some evaluation researchers, the act of evaluating can also be understood as a ritual, where the evaluand passes through a phase of liminality (Dahler-Larsen 2012). First, the evaluand is constructed and separated from everyday practice (pre-liminal phase). Then the process of evaluating creates a state of questioning, reflection and contestability, where the evaluand is caught "betwixt and between" (Turner 1995) the former practice and a new and indeterminate future practice (liminal phase). Finally, the end of the evaluation process reinstates the evaluand into existing structures and routines, but now in a more or less changed state – as a consequence of the evaluation (post-liminal phase).

To understand evaluations in the terms of liminality, is then to understand why the insertion of the evaluand in a liminal state is necessary, if evaluations are to succeed in transforming the evaluand. However, it also elucidates why this liminal state needs to be temporary. If the contestability of the evaluand is made permanent, the evaluand is placed in a permanent liminal state – i.e. being in a continual state of flux and unable to settle into a new role in a post-liminal phase.

In the example of the Danish employment system, such malleability of the evaluand is constantly upheld through the evidence-based evaluation system. The backbone of this evaluation system is a digital knowledge bank (on the domain jobeffekter.dk - jobeffects in English), where the Ministry of Employment gathers the available impact-evaluations of different employment programs and interventions – such as job training, educational courses or economic sanctions. Each of the collected impact-evaluations are given a value

according to their methods and data – the closer to the golden standard of the Randomised Controlled Trial (RCT), the better – and their results are synthesised in order to determine which programs work and which do not. This continual accumulation and publication of impact-evaluations is a way of permanently contesting the value of different employment programs and measures. Given that the municipal job centres are responsible for realising these programs, the knowledge bank is also contributing to permanently contesting the value of the job centres. This primarily happens in two ways.

Firstly, by continually evaluating existing or new employment programs in the job centres. Since 2005, the Ministry of Employment has regularly commissioned new RCTs - as well as quasi-experimental- and other types of impact-evaluations – on different job centre interventions. The job centres are thus well aware of how their practices are being continually evaluated by the Ministry of Employment and how this knowledge is fed into a larger evaluation system.

Secondly, the evaluation system is permanently contesting the value of the job centres by monitoring their compliance with the evidence-based knowledge of what works. As the Ministry of Employment is accumulating more and more impact-evaluations in their knowledge bank, their assuredness in this knowledge also grows. The evidence-based evaluation system is thus increasingly linked to a system of monitoring the activities of the job centres and whether these adhere to the prescriptions of the impact-evaluations. The logic being that enhanced job centre performance is assumed to naturally follow from the use of evidence-based programs. By failing to comply with the prescriptions of the high-ranked impact-evaluations, the job centres thus face the risk of unwanted scrutiny from the Ministry.

This permanent contestability renders the job centres highly malleable, but at the same time impedes their ability to change and develop. On the one hand, job centres are constantly adapting to the demands of the evaluation system - whether it be by implementing a new pilot-project according to the Ministry's detailed and highly specified project plan or by increasing the production of a given activation measure promoted by the Ministry. On the other hand, this constant flux makes it nearly impossible for the job centres to change their practice on a deeper level and with a clear aim. Robbed of any sure footing from which to set out on a new path, the job centre is reduced to either parroting the recommendations of the evaluation system or facing increased contestability. This is acknowledged by several of the interviewed job centre managers. They explain how the choice of job centre strategy is often based on the need to adhere to the recommendations of the Ministry of Employment, rather than to the perceived needs of the local context. This focus on compliance has been further bolstered in recent years, as job centres can now be placed under increased ministerial supervision – and thus intensified contestability - if their use of so-called evidence-based programs remain limited. In the words of the head of the employment services in one of the biggest municipalities in Denmark, this makes the job centres "world champions of implementation" (Interview on the 12/13/18), while leaving them severely lacking in innovation.

The permanent contestability differential created by the evaluation system infuses the Ministry with a steadfast conviction in the solidity of its own

knowledge-base, while simultaneously creating a permanent sense of doubt and lack in the evaluand. This ultimately restructures the focus of both. Evaluation becomes less of a tool for learning how to handle the societal problem, which the evaluand is trying to address – as for example unemployment – and more of a tool for fixing the evaluand's lack of compliance with the recommendations of the evaluation system. This situation resembles the "furious standstill" which Christina Segerholm – borrowing the term from Hartmut Rosa's theory of social acceleration – finds in her analysis of an evaluation system within Swedish higher education (Segerholm 2020, 621). The point being, that the permanent contesting of the evaluand leaves it in a fragile liminal state – always furiously and franticly adapting to the changes brought on by the evaluation system, but without any solid ground on which to stand and set out on a markedly new path.

The Paradox of Organisational Embeddedness

> *The organisational embeddedness of evaluation activities increases the reflectivity of the organisation, but decreases the self-reflectivity of the organisation*

One of the defining elements of evaluation systems is the way they embed procedures for commissioning, producing and using evaluations within public sector organisations. The traditional separation of the responsibility for the different parts of an evaluation process (from commissioning to use) is thus conflated as the same organisation can potentially function as commissioner, evaluator, evaluand and user of evaluations. This organizational embeddedness of evaluation activities is intended to build the evaluation capacity of the organisation and thus make it more apt at understanding evaluations and using their information correctly (Preskill and Boyle 2008) - ultimately creating a more knowledgeable and reflective organisation.

However, if we apply the concept of contestability differential, the build-in paradoxes of this organisational embeddedness become apparent. To be both evaluator and evaluand at the same time is to be simultaneously contested and uncontested, which would amount to trying to turn a screw without any solid grip or footing (Dahler-Larsen 2015a). Evaluations will thus always have a blind spot concerning their own perspective, as they cannot apply the same rigorous methods for reflection and questioning on themselves as they do on the evaluand. An evaluation can measure the efficiency of different activation programs and use this observation to distinguish between efficient and inefficient programs, but it cannot simultaneously question and observe the notion of efficiency/inefficiency from which it draws this distinction.

It may not seem terrible consequential or problematic that evaluators are unable to "take their own medicine" (Dahler-Larsen 2011). After all, evaluators are generally hired because of the quality and usability of their evaluations rather than the quality of their self-evaluations. However, this perspective is complicated when the evaluator is internal to the organisation responsible for solving the problems being evaluated. Then the lack of self-evaluation and self-reflectivity is, in fact, also a lack of reflection on the problem at hand.

In the case of the evidence-based evaluation system, the institutionalisation of procedures, routines and norms for developing and maintaining the system

within the Ministry of Employment, has created a peculiar form of double standard.

On the one hand, the civil servants of the Ministry's analytical departments routinely examine the existing knowledge-base in order to determine the need for conducting new (impact)evaluations. This continual reassessment and update of the knowledge-base of the evaluation system functions as a corrective measure aimed at the Ministry itself. Internal procedures have thus been established to make sure that the policies developed within other departments of the Ministry are checked by the analytical departments. Such procedures are meant to both discipline the other departments – i.e. make sure that they develop policies based on sound evidence - and alert the analytical departments of any need to produce new evidence.

On the other hand, this seemingly high degree of (self)reflectivity regarding the knowledge-base of the evaluation system, is countered by an unflickering trust in the soundness of this same knowledge, when it is disseminated to the job centres. The tentative, temporary and contextual conclusions of the existing knowledge-base are thus repackaged as uncontested evidence and then transferred into binding regulations on the job centres (such as the law stipulating the minimum number of caseworker-client meetings) or generic and universally applicable tools disseminated to the job centres (such as the method of Individual Placement and Support (IPS)).

This double-sided position of the Ministry of Employment is illustrated quite clearly in the interviews with high-ranking ministerial officials and evaluators responsible for conducting the earliest RCTs in middle of the 2000s. RCTs, which laid the groundwork for the later institutionalisation of the evidence-based evaluation system. When asked about the process of conducting these early RCTs, the interviewees highlight a host of methodological uncertainties related to their – hitherto - lack of experience in designing and implementing such experiments. These early studies are described as a shaky learning process rather than the solid foundation of the later evaluation system. However, when the same interviewees are asked to exemplify which evaluations have been the most influential on both the making and implementation of employment policies, they unanimously refer to these early RCTs. In the words of a ministerial top-manager, these early RCTs "created some main roads within the employment system, where the effects are well-documented" (Interview on the 04/18/18). Gone are then suddenly the beforementioned methodological caveats.

If we interpret these seemingly contradictory statements through the prism of the contestability differential, it becomes clear that the interviewees cannot truly question the now well-established truths of these older RCTs without also sacrificing the dominant knowledge position from which the Ministry currently exerts its influence. Just as the permanence of the evaluation system fixates the job centres in a fragile and constant state of flux (cf. the preceding section), the organisational embeddedness of the evaluation system also fixates the Ministry – albeit in a much more solid and firmly anchored position. The ultimate consequence, however, remains the same: the inability of the Ministry to set out on a markedly different path.

This fixation of the evaluator-role within the Ministry creates a rather powerful platform from which to influence the daily workings of the job centres, but it also renders the Ministry vulnerable when circumstances change rapidly. This is due to the fact that the Ministry of employment is ultimately part of the same chain of accountability as the job centres. The Ministry is therefore not only accountable to its own methodological standards – as external evaluators ideally are – but also to the will of the people (as represented by the elected politicians). The inability of the Ministry to step out of the evaluator-role and contest its own premise and perspective – due to the permanence and organisational embeddedness of this role – thus makes the ministerial civil service vulnerable to the shifting demands of governments.

Many of the interviewed civil servants of the Ministry of Employment exemplifies this by highlighting the change from bourgeois to centre-left government in 2011. The interviewees describe how the Ministerial civil service - prior to this shift in government – generally had come to the understanding that the employment-effects of educational courses were non-existing or even detrimental (Andersen 2020). However, this consensus was challenged by the new Social Democratic Minister of Employment, who wished to increase the use of this form of activation. This sparked a disruption within the Ministry, which the Minister has later described as a battle with the civil servants (Winther 2016) and the civil servants have described as a process of learning about their former blind spots (Andersen 2020). No matter which description most accurately fits the bill, the fact remained that the Ministry – in the eyes of the Minster – had hitherto failed to properly evaluate the effects of this specific type of activation. Ultimately, this hindered the Ministry from setting out on a different path in accordance with the Minister's political goals. This innateness is a direct consequence of the self-referential blind spot embedded into the organisation through the evaluation system. The Ministry of Employment is constructed as the firmly anchored subject who is permanently turning the screw, while being unable to observe and reflect on its own position – and thus remaining in the same place itself.

The Paradox of Epistemological Coherence

> *The epistemological coherence of the evaluation system increases the use of evaluations, but decreases the usability of evaluations*

The most unique aspect of an evaluation system is the shared epistemology that binds its different elements together. This epistemology clarifies: 1) what kind of knowledge the evaluation system seeks to produce; and 2) how this knowledge is best produced – i.e. through which evaluation methods (Leeuw and Furubo 2008). The primacy of this epistemological position is then upheld through the institutionalisation of specific standards and procedures for producing knowledge. Furthermore, such standards are also helping the evaluation system maintain a strong contestability differential between evaluation system and evaluand. Most evaluation systems – whether based on impact-evaluations, performance monitoring or audits – use quantitative methods and data not easily decipherable by laymen. These methods and data work on both sides of the evaluation-evaluand divide to bolster the contestability differential. The

evaluand is made more contestable by being quantified and made calculable and comparable (Dahler-Larsen 2019b), while the evaluation is solidified by being aligned with the dominant social imaginary, which values quantitative measures and methods above all else (Desrosières 1998; Porter 1996). A coherent and sophisticated set of methodological standards and guidelines are thus both a defining feature of evaluation systems and an important element in upholding a strong contestability differential. However, these same standards are also directly impeding the usability of evaluations.

To understand why this is the case, we can draw on a similar paradox well-documented within the bourgeoning literature on the adverse effects of performance indicators (e.g. Bevan and Hood 2006; Brodkin 2011; Dias and Maynard-Moody 2006; Fording, Schram, and Soss 2011; Munro 2004; van Thiel and Leeuw 2016). On the one hand, these – and similar – studies find performance indicators to be hugely influential on the behaviour of the organisations being measured. On the other hand, the studies also document how performance indicators often neither measure nor influence behaviour in the way they promise. This tendency is most clearly stated in the "laws" of Goodhart (Goodhart 1981) and Campbell (Campbell 1979), which argue that any indicator that becomes a target for social decision-making, is subject to corruption and thus ceases to be a good indicator. Herein lies a performance paradox, which can be stated in the following way: We get what we measure, but we cannot measure what we want to get.

A similar paradox is at play within the contestability differential upheld through an evaluation system. The power of – and adherence to – the information produced by the system hinges on it being produced through a set of coherent, standardised and highly sophisticated methods. However, as the programs evaluated by the system are implemented in differing and changing contexts, the betterment of these programs necessitates non-standardised and more methodologically flexible knowledge. We thus have a similar situation to the performance paradox, where the evaluand's adherence to the knowledge of the evaluation system hinges on the system's epistemological and methodological rigidity (you get what you measure), but the knowledge's usability for the evaluand hinges on its epistemological and methodological flexibility (you cannot measure, what you want to get).

In the case of the evidence-based evaluation system, the epistemological foundation is grounded in the standards of experimental and quasi-experimental impact-evaluations. Such evaluations are placed at the top of the Ministry's knowledge hierarchy and it is thus only knowledge produced through these methods, which is deemed evidence-based. However, the narrow methodological standards of the RCT-design also challenge the applicability of this knowledge to the diverse setting of 98 different municipal job centres. The question being, whether the specific intervention (such as job training or client-caseworker meetings), target group (the specified category of unemployed) and effect (for example a 10 pct. average decrease of the duration of the unemployment period) measured in a pilot experiment, can be replicated when these programs are rolled out to a larger population and context (Andersen & Randrup 2017).

This question was not given much attention in the initial strategy used for disseminating the findings of the RCTs during the 2000s and the beginning of

the 2010s. The knowledge was primarily transferred to the municipalities by being inscribed into the law and economic incentives regulating the job centres. The logic being, that the interventions tried in the RCTs were generic tools, which were therefore expected to generate the same positive effects - no matter the context or target-group to which they were applied. While this strategy has largely been effective in increasing the job centres use of the interventions deemed evidence-based, the job centres have not achieved the same employment effects as promised by the RCTs. Although the reasons for this are naturally manifold, the lack of similar employment effects has created doubts and concerns about the usability of the evidence-based knowledge on both sides of the evaluation-evaluand divide – albeit in markedly different ways.

Within the Ministry of Employment, the problem is conceived as primarily a problem of implementation. That is, the job centres lack the necessary knowledge to implement the evidence-based interventions correctly. The solution has been to both expand the knowledge-base of the evaluation system – for example by including qualitative indicators of why something works and how it can be implemented – as well as expanding the knowledge-dissemination strategy to also include "softer" measures than law-stipulated regulations. Among the most important of these softer measures is the translation of the evidence-based findings into ready-made "packages" containing the concrete tools and processes to be implemented in order to correctly apply the evidence-based interventions in the job centres (Nielsen, Danneris & Andersen 2020).

Within the municipal job centres, the doubts are more concerned with the general validity of the evidence-based knowledge. As a response, a large number of job centres have initiated so-called "investment projects" in order to develop their own knowledge of what works in their local context. The content of these investment projects – and whether it aligns with or deviates from the evidence-based prescriptions - differs significantly between municipalities. However, all the projects follow a similar process: First, a new type of intervention is introduced to a specific target group in a given period of time; then follows an evaluation of the outcome; and finally – on the basis of this evaluation - a decision is made on whether to terminate, continue or expand the intervention.

In different ways, both job centres and the Ministry are thus trying to broaden the scope of the evaluative knowledge, but as both strategies remain structured within the confines of the evaluation system, neither can escape the paradox of epistemological coherence.

Regarding the Ministry's strategy, the new types of qualitative and contextually sensitive evaluations are restricted to act as helping tools for the high-ranking impact-evaluations – rather than as a supplementary or competing knowledge-perspective. These evaluations thus have two main functions. The first is in the preliminary phase, where hypotheses are generated to later be tested through experimental methods. The second function is in the implementation-phase, where they are used to evaluate how the job centres implement the evidence-based interventions. In both cases, the problem is perceived as a problem of implementation, rather than having anything to do with the narrow conception of evidence-based knowledge within the Ministry.

A similar subsummation of alternative knowledge perspectives is taking place within the job centres, where the locally developed investment projects are challenged in two – related but different - ways.

Firstly, by the aforementioned pre-packaged implementation projects developed within the Ministry. These projects are promoted as a kind of investment project in their own right. The only difference being that they are financed by the Ministry rather than by the municipalities themselves. However, with this ministerial funding comes a fully-formed and detailed project-plan specifying the content of the given intervention as well as the process for implementing and evaluating it. Failure to adhere to such plans can lead to a withdrawal of the Ministry's funding.

Secondly, the municipally job centres are also placing themselves in a fragile position even when they choose to opt out of such pre-packaged projects. The job centres can then try out a completely new intervention, evaluate it themselves and maybe even succeed in achieving their goals. However, the job centres do not have the resources – e.g. manpower, expertise, access to data or a large enough population of participants - to evaluate their own interventions according to the (experimental) standards or the evaluation system. They are therefore unable to legitimise their own interventions and will – if these deviate from the evidence-based prescriptions – remain vulnerable to increased ministerial scrutiny.

Although the paradox of use vs. usability of the evaluation system is in a way acknowledged by both sides of the evaluation-evaluand divide, the dominant epistemological position – and the methodological rigidity following from this – remains unchallenged. The paradox thus remains, as the evaluation system continually upholds the notion of the evaluator as the knowledgeable teacher and the evaluand as the unknowing pupil. Never allowing the pupil to contest and adapt the teachings to its own needs, nor allowing the teacher to critically examine and develop its own teachings.

Discussion and Conclusion

The article has argued that evaluation systems simultaneously enhance and decrease the power of evaluations. This seemingly paradoxical situation is caused by the way evaluation systems make the contestability differential – and the asymmetric power relation herein - between evaluation and evaluand permanent, organisationally embedded and epistemologically fixed. Evaluation systems thus bind both the actors responsible for the evaluation system and the actors responsible for the evaluand to fixed and static subject-positions. On the one hand, this maximises compliance with the recommendations of evaluations - thus increasing evaluations' power over evaluands. On the other hand, this fixation of subject-positions and epistemological perspective also decreases evaluations' power to invoke radical change and development.

By theorising on this paradoxical nature of evaluation systems, this article thus contributes to our understanding of the power of evaluations in different ways.

Firstly, the findings complicate the picture – prevalent within the evaluation utilisation literature – of institutionalisation of evaluation activities being a

solution to solve the Gordian knot of non- or misuse of evaluations (Swee, Clark and Linda 2004; Läubli and Mayne 2016; Mayne 2009; Stockmann, Meyer and Taube 2020). If the goal of evaluations is to influence the evaluand – no matter the type of influence – then it is certainly true that the institutionalisation of evaluation systems can attribute to the achievement of this goal. There is, however, no guarantee that such influence will align with the types of use and influence typically promoted as desirable and intended by the evaluation utilisation literature – for example learning or enlightenment (Weiss 1998). To the contrary, the example of the evidence-based evaluation system in the Danish employment services suggests that evaluation systems can be detrimental to the ability to reflect and learn on both sides of the evaluation-evaluand divide – at least if these abilities are understood as more than just replicating the recommendations of evaluations (see for example Dreyfus and Dreyfus 1986). We should thus be wary of the idea that the greater institutionalisation and systematisation of evaluations, the greater the use (in both the descriptive and normative meaning of the word great) and focus on the actual consequences of such institutionalisation processes.

Secondly, this article also contributes to a recent sceptical turn within evaluation research (e.g. Dahler-Larsen 2019b; Furubo and Stame 2019; Segerholm et al. 2019). As explained by Dahler-Larsen, the increasing institutionalisation and systematisation of evaluation activities within modern societies have made some of the inherent problems of evaluations much more clearly visible – thus enabling us to view the act of evaluating with greater scepticism (Dahler-Larsen 2019b). Although the current article is clearly sympathetic to such a sceptical turn, the above findings also point towards a more radical departure from mainstream evaluation research. Rather than evaluation systems simply enhancing tendencies already at play within single evaluations, the inherent paradoxes suggest that evaluation systems may be an entirely different phenomenon than single evaluations.

If evaluations are always dependent on establishing a contestability differential by means external to the individual evaluation (be they regulatory, discursive or material), then the evaluation system functions as exactly such means. The creation of a strong contestability differential is thus not just an important factor for an evaluation system to function, it is the very essence of the evaluation system. It could therefore be analytically fruitful to study the construction of evaluation systems as an inherently different endeavour than single evaluations. An endeavour that has more to do with the governance, steering and management of public policies and programs than with "enlightenment" (Weiss 1998) and "speaking truth to power" (Wildavsky 1979). It is not that these aspects are necessarily mutually exclusive, they are just radically different and should be studied as such. The asymmetric power relation between evaluation system and its evaluand should therefore not be understood as an unintended side-effect, but rather as the very means through which the system makes the evaluand governable and manageable. This is of course especially evident in the current case, where the evaluation system is tied directly to the Ministry of Employment's supervision of the local job centres. Given that not all evaluation systems are this closely linked to the centre of power in an organisational field, the contestability differently between system

and evaluand would probably not be as strong in most cases. This is an important caveat regarding the generalisability of the current article's empirical findings. However, it does not necessarily affect the generalisability of the theoretical argument, as the embeddedness of an evaluation system within one or more of the central organisations in an organisational field remains a defining feature of such systems (cf. Leeuw & Furubo 2008). The evidence-based evaluation system of the Danish Ministry of Employment may not be representative of all evaluation systems, but it is certainly a highly successful one – when judged according to the internal logic of evaluation systems. This logic being, that the stronger institutionalisation of an evaluation system – on each of the four criteria proposed by Leeuw and Furubo (2008) - the more likely it is to enhance evaluation use. You can thus find many instances of less organisational embedded or less epistemological coherent evaluation systems, where the paradoxes would probably also be less vivid. But these more organisationally flexible and epistemologically reflective evaluation systems – if indeed they could still be defined as such – would then also be grounded in a weaker contestability differential and would therefore be less likely to increase evaluation use. The inherent paradoxes of evaluation systems thus remain: If such systems are to successfully achieve their stated purpose of increasing evaluation use, they need to create a strong contestability differential. However, if evaluation systems succeed in creating a strong contestability differential, they will also limit their likelihood of inducing actual change and learning within the evaluand.

The point is not to relieve evaluation systems of being judged according to ideals of learning and problem-solving – so often heralded as the guiding principles of evaluation activities. Nor is it to downplay the problems connected to the power asymmetries of such systems. Quite the contrary. By acknowledging evaluation systems as tools of governing, we are better able to critically examine the lofty promises and ideals employed by proponents of such systems. Furthermore, questions concerning the power relations and power effects of evaluation systems are brought to the fore of the study, rather than being placed at the margins – as is often the case within traditional evaluation research.

In the end, the analytical strength and promises of these theoretical arguments are, of course, for future research to judge. Moving forward, the focus on power asymmetries and paradoxes in other empirical cases of evaluation systems can help further our understanding of when and how systematization of evaluation activities turn into highly institutionalized and powerful evaluation systems. Thus, enabling us to not only view the systematisation of evaluations in a critical light, but also elucidate whether and how evaluation activities can be institutionalised in ways, where the benefits of evaluations can be reaped and the paradoxes of evaluations systems can be minimised.

Disclosure Statement

The author declared no potential conflicts of interest with respect to the research, authorship, and/ or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Alkin, Marvin C. & Jean A. King (2016) The historical development of evaluation use, *The American Journal of Evaluation*, 37(4): 568-579. doi:10.1177/1098214016665164

Alkin, Marvin C. & Jean A. King (2018) The centrality of use: Theories of evaluation use and influence and thoughts on the first 50 years of use research, *The American Journal of Evaluation*, 40(3): 431-458. doi:10.1177/1098214018796328

Andersen, Niklas A. (2020) The constitutive effects of evaluation systems: Lessons from the policymaking process of Danish Active Labour Market Policies, *Evaluation: The International Journal of Theory, Research and Practice*, 26(3): 257-274. doi-org.zorac.aub.aau.dk/10.1177/1356389019876661

Andersen, Niklas A., Dorte Caswell & Flemming Larsen (2017) A New Approach to Helping the Hard to Place Unemployed: The Promise of Developing New Knowledge in an Interactive and Collaborative Process, *European Journal of Social Security*, 19(4): 335-352. doi.org/10.1177/1388262717745193

Andersen, Niklas A. & Anders G. Randrup (2017) Evidensbaseret politikudvikling: Viden som oplysning eller indskrænkning af beskæftigelsespolitikken?, *Tidsskrift for Arbejdsliv*, 19(2): 41-56. doi.org/10.7146/tfa.v19i2.109071

Bevan, Gwyan. & Christopher Hood (2006) What's measured is what matters: Targets and gaming in the English public health care system, *Public Administration*, 84(3): 517-538. doi:10.1111/j.1467-9299.2006.00600.x

Bjørnholt, Bente & Flemming Larsen (2014) The politics of performance measurement: 'Evaluation use as mediator for politics', *Evaluation*, 20(4): 400-411. doi:10.1177/1356389014551485

Bourgeois, Isabelle & J. Bradley Cousins (2013) Understanding dimensions of organizational evaluation capacity, *American Journal of Evaluation*, 34(3): 299-319. doi:10.1177/1098214013477235

Brodkin, Evelyn Z. (2011) Policy work: Street-level organizations under new managerialism, *Journal of Public Administration Research and Theory*, 21(2): 253-277. doi:10.1093/jopart/muq093

Campbell, Donald T. (1979) Assessing the impact of planned social change, *Evaluation and Program Planning*, 2(1): 67-90. doi:10.1016/0149-7189(79)90048-x

Dahler-Larsen, Peter (2011) Taking One's Own Medicine? The Self-Evaluation of the Danish Evaluation Institute. In Pearl Eliadas; Jan-Eric Furubo & Steve Jacob (eds), *Evaluation: Seeking Truth or Power?* (pp. 75–88). Transaction Publishers, New Brunswick.

Dahler-Larsen, Peter (2012) The evaluation society, Stanford Business Books, Stanford.

Dahler-Larsen, Peter (2015a) The Evaluation Society: Critique, Contestability, and Skepticism, *Spazio Filosofico*, 1(13): 21–36.

Dahler-Larsen, Peter (2015b) Evaluation as Social Construction. In Thalia Dragonas; Kenneth J. Gergen; Sheila McNamee & Eleftheria Tseliou (eds), *Education as Social Construction: Contributions to Theory, Research and Practice* (pp. 315–335), Taos Institute Publications, Chagrin Falls.

Dahler-Larsen, Peter (2019a) *Quality: From Plato to performance*, Springer International Publishing AG, Cham. doi:10.1007/978-3-030-10392-7

Dahler-Larsen, Peter (2019b. The Skeptical Turn in Evaluation. In Jan-Eric Furubo & Nicoletta Stame (eds), *The evaluation enterprise – A critical View* (pp. 58–80), Routledge.

Dreyfus, Stuart E. & Hubert L. Dreyfus (1986) *Mind over Machine*, Free Press, New York.

Desrosières, Alain (1998) *The politics of large numbers: A history of statistical reasoning*, Harvard University Press, Cambridge.

Dias, Janice J. & Steven Maynard-Moody (2006) For-profit welfare: Contracts, conflicts, and the performance paradox, *Journal of Public Administration Research and Theory*, 17(2): 189-211. doi:10.1093/jopart/mul002

Flyvbjerg, Bent (2006) Five misunderstandings about case-study research, *Qualitative Inquiry*, 12(2): 219-245. doi:10.1177/1077800405284363

Fording, Richard; Sanford F. Schram & Joe Soss (2009) The organization of discipline: From performance management to perversity and punishment, *Journal of Public Administration Research and Theory*, 21(2): 203–232. doi.org/10.1093/jopart/muq095

Foucault, Michel (1982) The Subject and Power, *Critical Inquiry*, 8(4): 777-795.

Foucault, Michel (1990) *The History of Sexuality – Volume 1: The will to Knowledge*. Penguin Books, London.

Furubo, Jan-Eric (2019) Understanding the Evaluation enterprise. In Jan-Eric Furubo & Nicoletta Stame (eds), *The evaluation enterprise – A critical View* (pp. 3-31). Routledge.

Furubo, Jan-Eric; Ray C. Rist & Rolf Sandahl (eds.) (2002) *International Atlas of Evaluation*. Transaction Publishers.

Furubo, Jan-Eric & Nicoletta Stame (eds.) (2019) *The evaluation enterprise – A critical View*, Routledge.

Goodhart, Charles (1981) Problems of Monetary Management: The U.K. Experience. In Anthony S. Courakis (ed), *Inflation, Depression, and Economic Policy in the West*, Barnes and Noble Books, Totowa.

Hanberger, Anders, & Lena Lindgren (2019) Evaluation systems in local eldercare governance, *Journal of Social Work*, 19(2): 233-252. doi:10.1177/1468017318760788

Cousins, J. Bradley; Swee C. Goh; S. Clark & Linda, E. Lee (2004) Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge base, *Canadian Journal of Program Evaluation*, 19(2): 99-141.

Jacob, Steve; Sandra Speer & Jan-Eric Furubo (2015) The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later, *Evaluation*, 21(1): 6-31. doi:10.1177/1356389014564248

Läubli Loud, Marlène & John Mayne (2016) *Enhancing evaluation use: Insights from internal evaluation units*, SAGE, Los Angeles.

Lázaro, Blanca (2015) *Comparative study on the institutionalisation of evaluation in Europe and Latin America*, EUROsociAL Programme, Madrid.

Leeuw, Frans & Jan-Eric Furubo (2008) Evaluation systems – What are they and why study them?, *Evaluation*, 14(2): 157-169. doi:10.1177/1356389007087537

Luhmann, Niklas (1993) Observing re-entries, *Graduate Faculty Philosophy Journal*, 16(2): 485–98.

Mayne, John (2009) Building an evaluative culture: The key to effective evaluation and results management, *Canadian Journal of Program Evaluation*, 24(2): 1-30.

Munro, Eileen (2004) The impact of audit on social work practice, *The British Journal of Social Work*, 34(8): 1075–1095. doi.org/10.1093/bjsw/bch130

Nielsen, Mathias H., Sophie Danneris & Niklas A. Andersen (2020) The silent expansion of welfare to work policies: How policies are enhanced through the use of categorizations, evidence-based knowledge and self-governance. In Anja Eleveld, Thomas Kampen & Josien Arts (eds), *Welfare to Work in Contemporary European Welfare States: Legal, Sociological and Philosophical Perspectives on Justice and Domination* (pp. 163-188), Policy Pres, Bristol.

Nutley, Sandra, Isabel Walter & Huw T. O. Davies (2003), From knowing to doing: A framework for understanding the evidence-into-practice agenda, *Evaluation*, 9(2): 125-148. doi:10.1177/1356389003009002002

Nutley, Sandra, Anette Boaz; Huw T. O. Davies & Alec Fraser (2019) New development: What works now? Continuity and change in the use of evidence to improve public policy and service delivery*, Public Money & Management*, 39(1):1-7. DOI: 10.1080/09540962.2019.1598202

Oliver, Kathryn; Simon Innvar; Theo Lorenc; Jenny Woodman & James Thomas (2014) A systematic review of barriers to and facilitators of the use of evidence by policymakers, *BMC Health Services Research*, 14(1): 2. doi:10.1186/1472-6963-14-2

Porter, Theodore M. (1996) *Trust in numbers: The pursuit of objectivity in science and public life*, Princeton University Press, Princeton. doi:10.1515/9781400821617

Preskill, Hallie & Shanelle Boyle (2008) A multidisciplinary model of evaluation capacity building, *The American Journal of Evaluation*, 29(4): 443-459. doi:10.1177/1098214008324182

Rist, Ray C. & Nicoletta Stame (eds) (2006) *From studies to streams - Managing evaluative systems*, Routledge.

Rosenstein, Barbara (2015) *Status of National Evaluation Policies - Global Mapping Report (2nd Edition: February 2015)*

http://pfde.net/index.php/publications-resources/global-mapping-report2015

Scriven, Michael (ed) (1991) *Evaluation Thesaurus – Fourth Edition*, Sage Publications, Newbury Park.

Segerholm, Christina (2020) Evaluation systems and the pace of change - the example of Swedish higher education, *Educational Philosophy and Theory*, 52(6): 613-624. doi:10.1080/00131857.2019.1654372

Segerholm, Christina; Agneta Hult; Joakim Lindgren & Linda Ronnberg (eds) (2019) *The Governing Evaluation Knowledge Nexus - Swedish Higher Education as a Case*, Springer.

Sørensen, Eva & Jacob Torfing (2018) Governance on a bumpy road from enfant terrible

to mature paradigm, *Critical Policy Studies*, 12 (3): 350–359.

Stake, Robert E. (2003) Case studies. In Norman. K. Denzin & Yvonna. S. Lincoln (Eds.), *Strategies of qualitative inquiry* (2nd Ed.), (pp. 134 - 164), Sage Publications.

Stockmann, Reinhard; Wolfgang Meyer & Lena Taube (2020) The institutionalisation of evaluation in Europe. Springer International Publishing AG, Cham. doi:10.1007/978-3-030-32284-7

Turner, Victor (1995) *The Ritual Process: Structure and Anti-Structure*. Transaction Publishers, New Jersey.

Van Gennep, Arnold (2019) *The rites of passage (Second edition)*, The University of Chicago Press, Chicago.

van Thiel, Sandra & Frans Leeuw (2016) The performance paradox in the public sector, *Public Performance & Management Review*, 25(3): 267-281. doi:10.1080/15309576.2002.11643661

Vedung, Evert (1997) *Public policy and program evaluation*, Transaction Publishers, London.

Weiss, Carol H. (1998) Have we learned anything new about the use of evaluation?, *American Journal of Evaluation*, 19(1): 21-33.

Wildavsky, Aaron (1979) *Speaking truth to power – The art and Craft of Policy Analysis*, Transaction Publishers, New Jersey.

Winther, Bent (2016) *Mette F*, Berlingske, Copenhagen.