© *SpazioFilosofico* 2015 – ISSN: 2038-6788

Peter Dahler-Larsen

THE EVALUATION SOCIETY: CRITIQUE, CONTESTABILITY AND SKEPTICISM

Abstract

The essay begins with the observation that evaluation relates to its object in a manner that is not only descriptive, but rather constitutive. Five domains where the constitutive effects of evaluation occur are presented and illustrated. Next, three kinds of social critique are discussed, and counter-critique is offered. Each of these critiques is shown to coincide with particular ideas and roles such as "the authentic life before evaluation," "the rational architect of evaluation systems" or "the victim of evaluation." Finally, by using the concept of "contestability differential" as a can-opener, all evaluations are shown to rest on a combination of something which is contested with something which is not contested. On that basis, the essay concludes with a discussion of how a critique of evaluation can be cultivated in a democratic society.

1. *The evaluation wave*

Despite the inherent flaws in trying to reduce society to any *one* overarching dimension or principle (Morin 1988), the term *The Evaluation Society* does in fact capture many essential, frightening and awe-inspiring aspects of contemporary society. We do live in a society where evaluation, accreditation, auditing, benchmarking, performance management, quality assurance and similar documentation practices produce *datascapes* as an important dimension in social life along with idea-scapes, ethnoscapes, technoscapes etc. (Appadurai 1996). The function of these datascapes cannot be exhausted with reference to their descriptive aspects; instead it appears that they help define or constitute what they claim to measure. This observation raises the obvious question to which extent the contemporary evaluative grips on reality are conducive to structuring, regulating, and governing the social order in particular ways. Evaluating institutions may not be able to articulate how this takes place. So, the social, political and philosophical story-telling about evaluation should not be left to evaluators. What platform or position can be identified from which critique of the evaluation society can be articulated?

One reason why it is difficult to air critique is that evaluations are occupied with some large and positively sounding terms as *quality, sustainability, impact, equality, development, learning, transparency, innovation* etc. Since evaluators in their own view operate with indicators that approximately aspire to capture quality etc., evaluators often cannot understand why anybody would logically be against evaluation. Who are not in favor of quality? In contradistinction to earlier ideological tensions or class cleavages in society, the tensions around, say, quality, appear to be non-existent, because according to those in favor of evaluation, everybody must be able to get on board the mission for quality. If evaluation successfully captures all positive concepts little space is left for alternative views. The lack of recognition of conflictual material in the very ambition to achieve quality makes it difficult to argue that there even exists the possibility of a critical position.

It should also be noted that it is not without personal risk to seek such position. A story illustrates this. At a conference, a new bibliometric evaluation system for researchers was debated. A university lecturer aired a harsh critique of the attempts to measure quality of research through bibliometric indicators. The presenter at the conference session replied that in his view, there is a strong correlation between a researcher's bibliometric score and his or her general reputation. In other words, researchers with a good reputation have nothing to fear and have no particular reason to be critical. The breathtaking implication, never articulated, is of course that the critical academic was critical because he was not a good researcher. To immunize oneself against that kind of tacit accusations it would be necessary first to score well on bibliometric indicators and then prepare one's critique. Critics would thus have to work hard to earn the right to air their views. However, if they achieved good bibliometric scores, their motivation to undermine the trustworthiness and social acceptability of the score would be reduced. Perhaps one of the most important social logics of performance indicators is exactly this *divide et impera* between high-scoring and low-scoring members of the same group, regardless of the validity of the indicator.

In other words, in our attempt to articulate a position of critique and study what happens, we can learn quite a bit about how smartly and cleverly the evaluation society functions. The strategies it adopts in incorporating critique and fending it off may be quite advanced.

There is another reason why we need to consider critique of the evaluation society carefully. Critique often has a tacit normative component ("it would be much better if…"). Every critique identifies with some agent or position. It is important to be clear about these normative frameworks. If not, we run the risk of airing a critique that is not in sync with its own time and not sufficiently respectful of the subtleties of what it is critiquing.

Let me put this problem in another way. I teach students in political science. I teach them evaluation because evaluation is one of their functions in their future jobs. I also teach them to be critical of evaluation because I think it is an important socio-political phenomenon that no one should be blind or ignorant about. What do I expect of the critical views? That they are so special or so naïve or so normatively self-enclosed that they must be put aside when real evaluation is to be done… as if critics and evaluators have nothing to say to each other? Or do I think that the critical view is so advanced and so relevant that is must be taken seriously by evaluators, too? Truly, critique finds itself in a very ambiguous and delicate situation when it turns out that critique becomes useful in order to improve the social systems it criticizes, but contemporary capitalist and bureaucratic organizations have already for some time cleverly integrated various forms of critique into more optimal forms of system operations (Boltanski and Chaipello 2007). However, alternatively I would also be concerned if my students were evaluators until 5 p.m. and then critical after 5. p.m. The critical view should be aware of its own situatedness in society and acknowledge its own engagement in society.

My strategy in this paper is to begin, in an axiomatic way, with the illustration and depiction of an idea which I think is central to today's discussion about the evaluation society: The idea that evaluation stands not in a descriptive, but in a constitutive relation to what it claims to measure. Then I will subject this idea to three kinds of fairly conventional forms of critique, but what is new is that I will also be critical towards the critique. In the final section I will introduce the concept of "a contestability differential" as an ever-present element in evaluation and I will discuss, on this basis, whether it is possible to live with evaluation in a democratic way, or perhaps, even to cultivate the critique of evaluation and the democratic potential in evaluation at the same time.

2. *Constitutive consequences of evaluation*

Quantification begins with establishing the categories into which social phenomena are put so that they can be counted (Porter 1994). Statistical work not only reflects reality but establishes it by providing the players with a language to put reality on stage and act upon it (Desrosières 2002, p. 352).

Desrosières thus suggests that there is a constitutive element in the very act of measurement (putting reality on stage in a particular way) as well as in the subsequent act *upon* that construction of reality (which may make the construction more "real"). In social life, we can imagine that these two kinds of acts are difficult to separate (one is done with the other in mind), but we can also imagine that a particular evaluative staging of reality is not very successful if it does not lead to subsequent acts. If successful, however, evaluation can produce *constitutive consequences*.

To make this construct more visible and operational, I suggest we can observe these consequences in five domains. Under each domain I shall give examples of effects that are (perhaps) surprisingly concrete, while it should be remembered that this is exactly how such effects become embodied, in the concrete rather than in the abstract.

First, evaluation has an impact on the content on some work or practice. For example, testing in education is known to lead to "teaching-to-the-test". The test has a "wash-back" effect upon teaching itself, not only upon the organization of lessons, but also upon the curriculum and the choice of topics and materials (McNeil 2000).

Second, evaluation has an impact of timing of practices. Like budgets which are defined on a monthly or yearly basis and thus impact upon the timing of economic behavior, evaluation regimes impose their own rhythms on practices. For example, museums, schools, universities, hospitals and prisons have institutional dispositions for particular ways of anchoring themselves in time but their "goals" and "effects" become located in time in new ways according to how they are evaluated. Many evaluative issues are difficult to measure with validity, so time often becomes the universal currency in which quality and performance are expressed. How quickly? How often?

Third, evaluation has an impact on the (re-)configuration of social roles and identities. For example when students are asked to assess their degree of satisfaction with a teaching program, a new student role emulated on the basis of a consumerist approach to education emerges. Next, teachers teach in particular ways when they are subject to student satisfaction surveys. Different degrees of popularity among teachers may challenge teacher-to-teacher relations and put different teachers in different relations with their school managers and principals.

Evaluation thus suggest a set of interrelated viewpoints quite similar to what Marx called *Charaktermasken.* There is a structural basis for the kinds of masks or roles which individuals take on in the evaluation society, such as "producer", "consumer", "manager" etc. At the same time, *Charaktermasken* are indicative of some level of ambiguity in relations between roles (can we unmask?), some question of the cleverness with which masks are carried (how cleverly do you perform with your mask?), and some tension regarding how each individual negotiates the relation between role-playing and personality (to what extent should I see my evaluation results as something that characterize me personally?).

These three kinds of constitutive effects tend to be interconnected like words, timing, and roles in a drama. They enroll each other in a larger evaluative assemblage.

Therefore, fourth, constitutive effects of evaluation (of the three kinds above) tend to coalesce into a larger world view that provides a sort of integrated or mythical image of what is going on. For example, with bibliometric indicators of research, it is suggested that what is interesting about research is only a particular kind of output called publications. The different kinds of products are allocated different kinds of statistical weight (depending on reputation, "impact factor", etc.). On that basis synthesizing scores are developed. The overarching assumption that makes all this possible is that research should be understood as production. In a similar vein, an underlying assumption in PISA is that education is international competition. The meaningfulness of this idea is undergirded by an assisting myth which is that all countries have the same educational goals (Meyer 2008).

Fifth, the constitutive effects of evaluation extend to how we know, to our sources of knowledge. If an indicator has the implications suggested above, the meaning of an indicator changes when it is used as a part of an evaluation regime (Vulliamy and Webb 2001). When people change their interpretations and their actions as a result of the indicator, the indicator as a form of knowing is implied. I will now push this argument one step further and show that the same kind of argument also applies to other kinds of knowledge (officially regarded as knowledge or not) which are touched upon or enrolled by evaluation regimes. This is important because the richness of our insight into a particular phenomenon or practice under evaluation often depends on several kinds of knowledge.

Let me give two examples related to bibliometric indicators of research. A new bibliometric indicator has been defined in my country. All publications on a predefined list of journals and publishers (which does, regrettably, not include "Spazio Filosofico") are allocated a certain amount of points. With the help of academic committees, all publications are divided into two groups depending on their reputation. Only the best 20 % is allowed into the privileged group that get more points than the rest. The purpose of this differentiation is to prevent a situation in which all researchers just produce more publications of bad quality in order to score more points, an effect known from a study in Australia (Butler 2003).

A comparative score of sums of points broken down by institution goes into an algorithm that determines the allocation of research funds across institutions every year.

In principle, one of the main assumed advantages in bibliometrics is the objectivity of counting publications as opposed to the alleged subjective assessment in peer review. This particular advantage, however, is illusionary, because the bibliometric indicator is not independent from peer review but rather dependent on it. All publications in the bibliometric system are subject to some form of peer view in order to be categorized as respectable enough for being counted. As a consequence, editors of books will contact me and say things like: "Professor Dahler-Larsen, we really invite you to contribute a chapter to a new anthology. It is subject to peer review and you will earn bibliometric points for your contribution. However, given your experience in the field, I am sure that the peer review will not lead to a rejection of your wonderful contribution. So please accept our invitation."

Since the bibliometric indicator builds on peer review, it must be included by those who need to use bibliometric points in their negotiations with others. Whilst peer review in its classical meaning could assume both formative and summative functions, it is now deprived of the latter. The link or association between bibliometrics and peer review (or enrollment of the latter by the former, in Latour's terms) does not leave the latter unchanged. In my analysis: As a function of the bibliometric indicator, we are now less sure of what the peer review might mean than we were before, but we do depend on it for our indicator system to function.

A second example: I asked one of the architects of the bibliometrical system how he wanted to make sure that I in fact planned to aim at the most prestigious publications just because they gave more points than the other ones. I could devise a cool strategy to make a high number of points just by producing many low-ranking publications. He answered that I was welcome to do so, but he believed that I would be so sensitive to my colleagues' assessment of my work that it is in my own interest to make sure that the balance between high-ranking and low-ranking publications on my CV is not too skewed. I concluded that in order to not run amok, the bibliometric indicator still needed to be balanced with more conventional academic values. In other words, to work well, the indicator needs to prey on values that it does not itself embody. In a similar vein, we can imagine that other indicators in order to not produce totally anti-social behavior, still need that we know such things as norms, reputation, helpfulness, good practice etc. At the same time, it may also happen that a formal indicator tends to undermine or redefine the meaning of other forms of knowing that are embedded in other social norms and practices. We cannot just assume that collegial relations and professional conscience remain the same after a new evaluation regime is introduced which preys on but does not respect these other norms and forms of knowing.

Critique one: Evaluation is antithetical to authentic life

I shall now, as promised, discuss a number of critiques of the evaluation society that all respond, one way or another, to the observation that evaluation helps constitute something. The first of these forms of critique says that evaluation constructs artificial artefacts and is therefore antithetical to authentic life.

For example, using the practice of teaching and the categories of constitution mentioned above as an example, evaluation imposes a measurement regime which is against the very nature of teaching. Evaluation intervenes in the definition of content that would otherwise be chosen freely based on pedagogical considerations, evaluation intervenes in the spontaneous relations between teachers and students, evaluation imposes artificial time frames upon the teaching practice, evaluation confuses the reality of testing performance with real learning, and it undercuts the forms of knowing that springs from learning as an existential, relational and contextually embedded form of experience.

However, the problem with this line of reasoning is that it assumes an authentic and natural form of teaching and engagement with teaching that is ontologically prior to our knowledge-creation about teaching. If we argue that evaluation is against the very nature of teaching we tacitly assume that teaching springs out of nature and we thereby ignore the many investments human beings have made in different epochs in the phenomenon of teaching (socially constructed views of the human child; the role of authority; the role of education in relation to society; the changing visions of the good society to which good education is a preparation; etc.).

In a broader perspective, the critique that says evaluation is antithetical to authentic life tends to assume a certain pre-social destiny handed down to us. It is the identification with this pre-given order of things that allows the critique of evaluation to point to the *artificial* nature of data. This view risks lending itself to uncomfortable subscriptions to a metaphysical order of life. It too easily allies itself with a traditional, religious or even totalitarian undertone. We *know* what the authentic life is, and it commands us to live in a particular way.

However, if you listen to such commands, you can hear them in many variations, sending you off in different directions. How far should we go back? If books and newspapers and diaries are tools for systematic reflection, should they also be abandoned? Should we abandon thermometers and ask ourselves if we feel warm? Should we live like the Amish? Should we break all mirrors because they allow us to see ourselves from the outside? Or is it OK to make a systematic data-based inquiry into the effects of tobacco on lung cancer, but not OK if we call it an evaluation?

The command that sends us back to an “authentic” form of life must ignore, in Cornelius Castoriadis (1997) view, the responsibility we have as modern human beings to organize our own world and make our own laws. It also ignores, I believe, Gianni Vattimo’s (2004) observation that if we “take on” and “work through” the contemporary socio-philosophical condition, we cannot operate with a “handed-down” or metaphysical guarantee to support any argument.

However, a modified and humble or “weak” variation of the argument is possible. It goes like this. It is not possible to be reflexive about everything. No social system can question all its operations at the same time. There simply is no capacity for that (Bateson 1972). If the evaluation society promises endless development, endless change, endless accountability, and endless reflexivity, it is giving us illusions. In fact, quite a lot of social critique in recent years says that the ever-performing subject is in fact presently at war with itself (Han 2012). Too much flexibility can be destructive of the social fabric of norms and of personal values (Sennett 2002), and it is possible to recommend a certain personal and social solidity that resists endless re-definition (Brinkmann 2014).

Evaluation transports a modern technical mentality according to which life consists of components that can be measured and replaced (Berger, Berger and Kellner 1973). Truly, there are forms of life that are embedded in frames of normativity that cannot be subject to any kind of evaluative perspective, any kind of componentiality. I am thinking of care, love, memory, pride, self-respect, geniality etc.

Modern existence seems to be caught in a paradox. Once we have discovered the reflexive standpoint, it is difficult to live as if the spontaneous form of life is the only one possible. We know it is not. We also know that there are existential "choices" or "ways of being" that lose their meaning if they are subjected to any kind of evaluative perspective. On a scale from one to ten, how do you assess the love of each of your children? There is great paradox in the fact that it requires (some kind of) reflexivity to even choose to protect such forms of life from (some kind of) reflexivity. It is difficult to *choose* to live spontaneously and authentically. It is like using one compartment of life to protect another compartment of life without succumbing to the compartmentalization of life.

Nevertheless, contemporary contributions to a critique of endless reflexivity seem to suggest that we doom ourselves if we have no "brakes" on mechanisms that enhance reflexivity. So, in a revised and moderated form, this first critique suggests that there is something which perhaps should be protected from evaluation not because it is authentic but because we care and find it wise to protect it.

Critique two: Evaluation has counter-intentional side effects

This critique often takes a starting point in the observation that measurement of complex phenomena is bound to be imperfect. Thus, indicators of the quality of public services, the impact of research, the innovation in the public sector, and sustainability of climate policies etc. are marred by flawed validity. Nevertheless, in a managerial context, these indicators are used for all sorts of purposes anyway (accountability purposes, steering purposes, information purposes etc.).

When imperfect measures are used, evaluation often has unintended consequences. So, according to this kind of critique, the problem with the constitutive consequences of evaluation is that they are *unintended* constitutive consequences. For example, if we measure the time from patients arrive at the emergency room until they encounter a nurse, some hospitals hire nurses to immediately say "hello" to each patient. The world is full of examples in which you can live up to what is being measured without living up to the intention behind the measurement. The discrepancy between the two is rooted in the validity problem described above.

For that reason, some advocates of evaluation spend quite a lot of time refining and cultivating the indicators used in evaluation, a process called purification by Latour (2004).

It can also be recommended to use a broader set of indicators (because there is an evaluation deficit in what is not being measured so far) or to use a more narrow set of indicators (because the general purpose has been lost in a jungle of measurements). There is a whole range of evaluative techniques concerning who gets measured how and when all of which can be varied in order to improve validity. For example, the measurement of *effects* is almost like a whole discipline in itself that includes various schools of thought.

What remains, however, is that as long as measurements are *approximations* to the perfect measure, there will be unintended consequences of evaluation in practical use. This idea is not extremely radical, because it is accepted as a sensible middle ground between strong critics and strong believers in evaluation, performance measurement etc. (Norman 2002). The key point is perhaps only whether there are so many and so important unintended consequences of evaluation that they constitute a substantial reason for objection, and not least importantly, whether these unintended consequences can somehow be repaired.

What I would like to stress here, however, is the underlying identification of that kind of critique with the idea of *intentions* in evaluation. Logically, unintended presuppose an intention on the other side of the conceptual coin. However, a number of questions can be asked here (Dahler-Larsen 2014).

How can intentions be captured empirically if they are not stated? Do not tell me we can trust official political declarations of intentions! Which intentions count? Do we imagine an architect behind the evaluation whose intentions we share? Could other players have intentions, too, and what if all these intentions are not in alignment? If people invent new intentions, do we then go back to some "original" intentions or do we allow people to invent intentions along the way? Do we assume intentions behind a particular indicator, evaluation, or evaluation system? What if a network of evaluative phenomena amounts to a whole surveillant assemblage (Haggerty and Ericson 2000)? Is it not meaningless to assume one set of intentions behind such dynamic network?

The critique that claims that evaluation has unintended consequences more often than not *identifies with an icon of an evaluation architect* that rationally seeks to plan and control evaluation with the best of all intentions, but, alas, unfortunately, misses the target because the indicators fail to support him all the way. Would it really be better if evaluation was planned and controlled all the way? And perhaps even more importantly: Why should an analytical perspective identify with the so-called architect of evaluation when there are so many other perspectives in society one can identify with? If a scientific perspective is one that does not identify with any particular part in a political situation, why should evaluation research identify with this imaginary and overly rational evaluation architect?

Why miss the evolving and dynamic character of spontaneous evaluative initiatives? Some of these initiatives may have constitutive effects that are, in fact, not counter-intentional, but rather quite consistent with *some* political intentions (such as the redefinition of content, the reconfiguration of social relations in the direction of something more flexible, componential, and marketable). But my counter-critique goes one step further.

If we acknowledge that statistics are constitutive of what they claim to measure, and we apply the intended/unintended distinction thereto, perhaps we too early curse a measurement because it was not agreed upon or it was not collectively intended rather than in fact study and understand how it, for better or for worse, feeds into our collective sense-making and society-building. In a democratic deliberative perspective, for instance, would we accept that some say "my values would support proposal A", and others would say "I fear that proposal B would be disadvantageous for the weakest members of our society", but not accept if one said "I have done a survey that leads to the conclusion that proposal C is the best proposal"?

In fact there exist some areas of political contention where the very ambition to do research or evaluation helps constitute that area as one that deserves attention. Some measurements of risks qualify here (Beck 1992). The same is true with the whole area of the "psycho-social work environment".

One might argue that numbers are used strategically to make an argument more objective or technical than it deserves to be, because it is really just a statement from a particular viewpoint. However, if we insert into our common deliberations the no-longer-radical idea that statistics are social constructions, too, we can acknowledge their pragmatic and socially constructive qualities without succumbing to them as if they were cast in stone. Numbers can fool us in a thousand ways (!). But it is also part of the history of numbers (e.g., as embodied in the metric system) that they are supposed to help us agree to some common understanding of some aspect of something (Porter 1995).

The intended/unintended distinction is not one that deserves to be applied routinely to the constitutive aspects of evaluation, as if it leads to the highest wisdom of all to know whether a phenomenon that happens is or is not in alignment with some reconstruction of some alleged original intentions.

Critique three: Evaluation is power

According to this third kind of critique, evaluation cannot be understood apart from its specific historical and institutional embeddedness. Many have found Orwell's "Big Brother" and Foucault's panopticon to been prime metaphors for understanding how the evaluation society combines surveillance with a structuration of the modern social order. Foucault's contribution is to highlight how techniques for measurement, documentation and comparison become practices for governing at a distance in way that also involves discipline and self-scrutiny of subjects (in the interesting double meaning of "subject to" and "subject for").

In education and in other fields there is a rich literature on colonizing evaluation practices that refer to Foucault (Shore and Wright 1999). Although it is probably correct in pointing to the link between evaluation practices and the larger institutional order, as well as to the production of monitorable subjects, this paradigm perhaps assumes too much of a centrally located point of observation, too much of a one-directional observation, and too much certainty about what is produced of what we have called "constitutive effects." I am reminded of Zizek's provocative warning that if we say that the outcome of totalitarianism is determinately known and nothing but tragic, we are almost giving the totalitarians too much. The best key to understanding these analyses, I think, is that they tacitly *identify with the victims of evaluation*. As if this category of victim is analytically easy to define, as if the members of this category are defined through and through by the "character mask" they wear, and as if the strategic move of victimization in itself supplies members of this category with some moral superiority. And as if the analysis of victimization takes place in a totally different world from the one in which victimization takes place. If the analysis of victimization is correct, how is there even space for a critical analysis? Maybe these questions are not asked because if victims are morally superior, it would not be a good idea to search for alternative positions. It is better to remain a victim. The clearer the power structure, the easier the identification with victims.

An attempt to paradigmatically update the surveillant assemblage in a more "undeterminate" direction is provided by Haggerty and Ericson (2000). They assume, with Latour, that there are scattered centers of calculations that are not necessarily hierarchically related (sometimes police is filming a demonstration, but activists also film the police). Some of the new technologies of documentation and registration (cameras, survey software) are inexpensive and dispersed in ways that do not conform to authoritarian hierarchies. There is instead "a highly fractured rhizomatic criss-crossing of the gaze such that no major population groups stand irrefutably above or outside of the surveillant assemblage" (Haggerty and Ericson 2000, p. 618).

There is a potentiality in surveillant assemblages that becomes actualized only in particular ways when particular connections are made. There are constant negotiations going on, and new connections lead to the invention of new uses, and sometimes "endless redefinitions and reconfigurations" (Callon 2010, p. 165). (I am not sure that Foucault would object to observations like these; what I note, however, is that some of his epigons do not take up that research agenda).

When risk is imposed upon a part of a political steering system, that part is like to push back in order to avoid the risk, which leads to "spiraling regulatory logics" (Rothstein, Huber and Gaskell 2006).

Thévenot and the "pragmatic sociology" (Boltanski and Chiapello 2007) take the discussion of Foucault in a slightly different direction. They argue that a Foucauldian world is unlivable. They believe, with Durkheim, that any society needs some sort of moral fabric. In modernity, we have a high number of moral repertoires to draw from in our construction of institutionalized solutions to common problems that are seen to be more or less legitimate. In other words, as a corollary, evaluation practices need some form of justification which can, in principle, be interactively debated.

For example, in a case study in Denmark, I followed upper secondary school teachers who were discussing the meaning and consequences of student satisfaction surveys in their schools. When doing so, they drew on different repertoires. A part of the discussion had to do with whether student satisfaction data are truthful, valid and reliable. Another aspect dealt with fairness and justice, for example whether it is fair to compare schools in different socio-economic districts and whether it would have been more fair to include teachers in the planning of the survey at an earlier stage. It was also discussed, at the same time, whether the student satisfaction surveys could be useful for improving the student climate at the school. Truthfulness, fairness and utility all served as registers from which to draw arguments.

Such an analysis perhaps focuses too much at the micropolitics of evaluation inside the upper secondary school, but I admit that micropolitics are connected to macropolitics: the broader education policy, marketization of schools, increased competition among schools, etc. All I am suggesting is that in a particular case, it may be worth looking not only at how evaluation supports one-sided and hierarchical power structures, but also how evaluation connects with actual arguments in a fragmented, diverse, and dynamic structure of power, including local negotiations.

Perhaps an a priori and general theoretical commitment to *either* a hierarchical power structure *or* a more flexible, diverse, fragmented and reflexive social order as conflicting paradigms is misplaced. Perhaps any particular socio-historical situation and any particular case study present us with a unique configuration that may draw differentially on the two paradigms, respectively.

3. *The contestability differential*

In a world handed to us by God or by tradition, evaluation cannot be carried out. Evaluation assumes that some aspect of social life is contingent. Evaluation is a planned inquiry deliberately designed to induce contingency. Evaluation assumes a set of expectations about potential social change, much like concepts in modernity open up a new horizon of expectations (Koselleck 2007).

Evaluation challenges a particular aspect of social life by saying: I will measure your quality, and maybe you need to change in order to improve what I define as your quality. It is the job of evaluation to make the evaluand contestable. It is easier to make sure evaluation is used if there is conflict and a pressure to act in the evaluation situation (Lederman 2012) which is equivalent to saying the evaluand is contested.

Evaluation is a special kind of social/institutional initiative because it is a practice that is deliberately organized in order to change another practice. To do so effectively, evaluation must protect itself from contestability. Evaluation needs to be backed up by, say, belief in methodology and data, in the credibility of the institution that carries out evaluation, and in the virtues related to using evaluation for good purposes such as learning or improvement. If evaluating institutions cannot count on such beliefs, they must have the power to carry evaluation anyhow. Without any of these social anchors, an evaluation would be futile.

A metaphor: 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. The person needs to make sure that his weight makes his feet stand solidly on the ground as he turns the screw. He or she also must make sure that the screwdriver has a solid grip on the screw. A child may not be able to do it. Perhaps it takes several attempts from a strong and heavy person with skills.

The same with evaluation. To function effectively, an evaluation must exploit the differential between the (relative) fluidity of the social material it seeks to change and the (relative) solidity of its own fixation in the world. I call this difference "the contestability differential." All evaluation plays with the difference between what is solid and what is not solid.

Alternatively, when a contestability differential cannot be established, evaluation cannot take place. We may have so much strength and power in traditions and in institutions that they cannot be evaluated (that is why we do not evaluate flags or royal families or the best and worst wars). OR: An evaluation is criticized so much for corrupt indicators, a filthy evaluation process, a manipulated result, or a lack of independence from political interest that perhaps there is not enough justification for using it. These observations, too, correspond to a failure in establishing a contestability differential: If the evaluation becomes more contested than what it seeks to evaluate, then it cannot operate.

Sociologically speaking, evaluation is a modern phenomenon that thrives on reflexivity and contingency. Evaluation makes its object soft and contestable and fluid. On the other hand, any particular evaluation itself needs to be relatively firmly anchored in something that is more solid and less contested. Evaluation can take on the "taken-for-granted" character that constructivists (Berger and Luckmann 1967) and institutionalists (Scott 1995) talk about. Evaluation can find support in normative, cognitive and regulatory institutional pillars such as belief in data-based decision-making, or incentives based on evaluation results. In Latourian language, we can talk about so many solid associations with various actants (people, inscription devices, resources, sanctions etc.) that it becomes possible for evaluation to operate as a "black box" that can be inserted as an operative element in large networks of activity.

The advantage of the contestability differential as a concept is that it allows us to see evaluation as a powerful force that (like the market) has the restructuring of social orders and relations as a primary function, without conceptually committing ourselves to always seeing evaluation fixed in the same way to any particular ideology or institution. Evaluation lends itself to more than one ideological agenda (Kipnis 2008). Neither are we committed to assuming that evaluation works deterministically in every instance. To work as a construction, it must first be constructed.

How does this take place in practice? It is necessary to ask this question because our belief in the value of the concept of contestability differential is sustained if the concept can be operationalized and used in empirical analysis.

Several options are available. For example, the evaluand (the object of evaluation) can be criticized for lack of effectiveness, quality etc. This seems to be one of the strategies that politicians use against public institutions such as schools. A softening of the object always makes evaluation easier. Next, evaluators and managers can talk smoothly about the many good consequences of evaluation (learning and development). They can align themselves with powerful institutional forces (expertise, manpower, management, financial incentives, legal consequences) and weave evaluation into organizational structures and processes through scripts and recipes such as "evaluation cultures", "evaluation capacity", "evaluation policies", and a "general need to be learning-oriented and flexible". They can also connect evaluation with evaluation imaginary (Schwandt 2009) in the larger social environment such as the myth of development or the myth of assurance (Dahler-Larsen 2012; Power 1997), the latter assuming that evaluation is the best response to a cultural anxiety about risk, crisis and potential disaster.

When evaluation is in a very powerful position, it does no longer need to justify itself (thus the deteriorating influence of "evaluability assessment", an old-school procedure in evaluation which served to check whether a potential evaluand was in fact ready to be evaluated) (Dahler-Larsen 2014).

In that case, the contestability differential works very well. The abolishment of evaluability assessment indicates that the belief in systematic evaluation has become so strong that evaluation does not need to justify itself in each and every instance.

However, it may be costly to establish a strong contestability differential. It is always a delicate matter how much resources and how much institutional power should be invested in systematic evaluation. Evaluation may be expensive, and evaluation based on institutional force is only complied with as long as the subjects are faced with sanctions and a sense of necessity. It is difficult managerial balance to achieve the benefits of soft control while only reverting to harder forms when necessary. In some situations, there is a struggle between different elements of varying degrees of contestability, and evaluation has to fight from house to house.

In contemporary evaluation there is sometimes a structure, a function or an organization that serves as an "evaluation machine" without any subjective or human representation. When we are scared about or worried about "evaluation", we are in fact faced with a large network of institutional elements, people, inscription devices, and resources that enroll us as "actants" with a particular "character mask" in relation to evaluation. We cannot always see the whole hinterland behind this construction. Nor are we interested. There is a "metadata" paradox here (Desrosières 2002). Although, from a technical or methodological perspective, we are interested in all the factors that influence on how evaluation results are produced, in a practical sense, we are not. We would be tired or die of boredom (Lindeberg 2007) or react much too slowly if we were to appreciate and understand all the details necessary to produce the large-scale evaluations; what we are faced with is that they are *actionable* already.

As analysts, however, it is our duty to tell a longer story.

4. *Evaluation and democracy*

If we recognize that all evaluation is built on some manifestation of a contestability differential, there is no universal normative prescription that commands us to identify with a particular "character mask" a priori or with a particular foundational principle or myth that structures evaluation. Instead, we should be skeptical about general standpoints and general identifications. Personally, I am very skeptical of aligning evaluation with too much power, i.e. too solid a contestability differential. I am uncomfortable with the usurpation of political power and democratic roles by evaluating institutions (Neave 1997). In my personal view, there is more need than ever before to debunk the way that evaluating institutions build up their contestability differentials through a variety of means. I am particularly skeptical about the automatization, institutionalization and standardization of evaluation as it takes place in alliance with powerful organizations, and I am skeptical of the link between evaluation and ideologies such as a neo-liberal idea of all-encompassing productivity, marketization and competitiveness.

At the same time, I am also skeptical about generally not being willing to build any contestability differential that would make evaluation possible. We can learn from the sociology of knowledge that all knowledge is due to some element of social construction, and some element of “black boxing” of what we think we know, but nevertheless we are doomed to live in a world where we must responsibly construct knowledge. Knowledge is a capacity to act (Stehr 2001). We are also doomed to find out how we can best handle our common social and political destiny through democracy. If politics and democracy are those domains where society works upon itself (Rosanvallon 2009), there is no need to totally abandon systematic and deliberate knowledge-production, although we have, of course, learned that knowledge production is not just descriptive but also constitutive. I do acknowledge that the “we” included in the previous sentence is also potentially contested. In Rosanvallon’s perspective, democracy is always historical and situational; we may later learn to discredit principles that served us well in an earlier epoch. Our democratic knowledge production by definition has a preliminary character.

Faced with that, it is democratically possible to ask for more evaluation in one area (wars) and less evaluation in others. I also think it would be fair to argue that some aspects of life are better protected without evaluation, although we may later change our priorities.

In a democratic context, it is a difficult task to build exactly the kind of contestability differential needed for a particular form of evaluation in a particular situation, not more, not less. We have to acknowledge the paradox inherent in this endeavor. We know we build that which we take for granted for a while. Everything can be contested, but not much at the same time. We know we have purposes, but our instruments do more than just help us with fulfilling these purposes in a transparent way. If evaluation is constitutive, it is by definition infused with ambiguity (Best 2008). There are no instruments that constitute things in a pure way. There is always an overflow. The good news is that once we have discovered the idea of the contestability differential as an ever-present ingredient in evaluation, we seek to provide any contestability differential with only the preliminary, temporary and fragile status it deserves.

I cannot be much more precise, but it is this kind of skeptical thinking about evaluation that I deem consistent with a democracy in which humans know that they set their own laws and live with the consequences. I think it would also be consistent with a kind of weak thinking (Vattimo 2004) according to which we construct our collective arguments paradoxically acknowledging that there are no firm foundations or guarantees undergirding our arguments about what we think we know. Evaluation can be based on no general and winning argument about truth, fairness or utility. We have to dare to make the humble and situated arguments as we go along.

References

- A. APPADURAI (1996), *Modernity at Large. Cultural Dimensions of Globalization*, University of Minnesota Press, Minneapolis MI 1996.
- G. BATESON (1972), *Steps to an Ecology of Mind*, Ballantine Books, New York NY 1972.
- U. BECK (1992), *The Risk Society*, Sage, London 1992.

- P.L. BERGER-B. BERGER-H. KELLNER (1973), *The Homeless Mind. Modernization and Consciousness*, Vintage Books, New York NY 1973.

- P.L. BERGER-T. LUCKMANN (1967), *The Social Construction of Reality*, Doubleday, New York NY 1967.

- J. BEST (2008), *Ambiguity, Uncertainty, and Risk: Rethinking Indeterminacy*, in "International Political Sociology", 2 (2008), pp. 355-374.

- L. BOLTANSKI-E. CHIAPELLO (2007), *The New Spirit of Capitalism*, Verso, London 2007.

- S. BRINKMANN (2014), *Stå fast - et opgør med tidens udviklingstrang*, Gyldendal, København 2014.

- L. BUTLER (2003), *Explaining Australia's Increased Share of ISI Publications. The Effects of a Funding Formular Based on Publication Counts*, in "Research Policy", 34 (2003), pp. 565-574.

- M. CALLON (2010), *Performativity, Misfires, and Politics*, in "Journal of Cultural Economy", 3 (2/2010), pp. 163-169.

- C. CASTORIADIS (1997), *World In Fragments. Writings on Politics, Society, Psychoanalysis, and the Imagination*, Stanford University Press, Palo Alto CA 1997.

- P. DAHLER-LARSEN (2012), *The Evaluation Society*, Stanford University Press, Palo Alto CA 2012.

- P. DAHLER-LARSEN (2014), *Constitutive Effects of Performance Indicators: Getting Beyond Unintended Consequences*, in "Public Management Review", 16 (7/2014), pp. 969-986.

- A. DESROSIÈRES (2001), *How Real Are Statistics? Four Possible Attitudes*, in "Social Research", 68 (2/2001), pp. 339-355.

- K.D. HAGGERTY-R.V. ERICSON (2000), *The Survelliant Assemblage*, in "British Journal of Sociology", 51 (4/2000), pp. 605-622.

- B.-C. HAN (2012), *Træthedssamfundet*, Møller Forlag, København 2012.

- A.B. KIPNIS (2008), *Audit Cultures: Neoliberal Governmentality, Socialist Legacy, or Technologies of Governing?*, in "American Ethnologist", 35 (2/2008), pp. 275-289.

- R. KOSELLECK (2007), *Begreber, tid og erfaring*, Hans Reitzels Forlag, København 2007.

- B. LATOUR (2004), *Why Has Critique Run Out of Steam? From Matters of Fact to Matters of Concern*, in "Critical Inquiry", 30 (2/2004).

- S. LEDERMAN (2012), *Exploring the Necessary Conditions for Use in Program Change*, in "American Journal of Evaluation", 33 (2/2012), pp. 159-178.

- T. LINDEBERG (2007), *Evaluative Technologies: Quality and the Multiplicity of Performance*, Copenhagen Business School, København 2007.

- G NEAVE (1998), *The Evaluative State Reconsidered*, in "European Journal of Education", 33 (3/1998), pp. 265-284.

- R. NORMAN (2002), *Managing Through Measurement or Meaning? Lessons from Experience with New Zealand's Public Sector Performance System*, in "International Review of Administrative Sciences", 68 (2002), pp. 619-628.

- T.M. PORTER (1994), *Making Things Quantitative*, in "Science in Context", 7 (3/1994), pp. 389-407.

- M. POWER (1997), *From Risk Society to Audit Society*, in "Soziale Systeme", 3 (1997), pp. 3-21.

- H. ROTHSTEIN-M. HUBER-G. GASKELL (2006), *A Theory of Risk Colonization: The Spiralling Regulatory Logics of Societal and Institutional Risk*, in "Economy and Society", 35 (1/2006), pp. 91-112.

- W.R. SCOTT (1995), *Institutions and Organizations*, Sage, Thousand Oaks CA 1995.
- R. SENNETT (2002), *The Corrosion of Character. The Personal Consequences of Work in the New Capitalism*, W. W. Norton & Company, New York NY 2002.
- C. SHORE-S. WRIGHT (1999), *Audit Culture and Anthropology: Neo-Liberalism in British Higher Education*, in "The Journal of the Royal Anthropological Institute", 5 (4/1999), pp. 557-575.
- N. STEHR (2001), *The Fragility of Modern Societies. Knowledge and Risk in the Information Age*, Sage, London 2001.
- G. VATTIMO (2004), *Nihilism and Emancipation: Ethics, Politics, & Law*, Columbia University Press, New York NY 2004.