Article

Evaluation
1–17

DOI: 10.1177/1356389018819110
journals.sagepub.com/home/evi

Evaluation culture and good governance: Is there a link?

Peter Dahler-Larsen
University of Copenhagen, Denmark

Adiilah Boodhoo
University of Cape Town, South Africa

Abstract

An important motivation for the institutionalization of evaluation culture in countries around the world is the belief that accountability and transparency will thereby be enhanced. We subject this narrative about evaluation's contribution to good governance to empirical analysis. We also argue that the meaning and relevance of this general narrative differs across national contexts. We build on data from a systematic assessment of evaluation culture in 19 countries (Jacob et al., 2015), add one country using the same systematic approach, and combine these findings with an indicator of transparency in government provided by Transparency International. We find a positive correlation between evaluation culture and transparency, and discuss threats to a causal interpretation hereof. We go into depth with two particular countries at either end of the transparency scale. We argue that the meaning of the link between evaluation culture and transparency differs whether the chosen perspective is generally comparative or situated in particular national, historical, and political contexts. In countries where transparency is high already, there might be diminishing marginal returns on evaluation, at least regarding its contribution to accountability and transparency.

Keywords

accountability, evaluation culture, good governance, transparency

Introduction

Countries all over the world are adopting values such as professionalism and transparency (Donaldson and Lipsey, 2006: 56). There is considerable interest in supporting and enhancing evaluation culture. An important motivation for the institutionalization of evaluation is that evaluation promotes accountability, transparency, and good governance. This assumption is

Corresponding author:
Peter Dahler-Larsen, Department of Political Science, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark.
Email: pdl@ifs.ku.dk

near to the hearts of many evaluators and resonates with how textbooks on evaluation present accountability as one of the key purposes of evaluation together with learning, knowledge-building, and informing the public (Chelimsky, 2006; Vedung, 1997). Consistent with sound evaluative thinking, the institutionalization of evaluation should not be seen as a purpose in and of itself. It should be justified in terms of its outcomes. The link between evaluation culture and transparency is a story about the larger almost spiritual mission of evaluation, but it is also a story which can be subjected to empirical testing.

It is a massive undertaking to collect valid and reliable comparable data about national evaluation cultures. We acknowledge the contribution of Jacob et al. (2015) but take their findings further by asking how they might be linked to good governance. We discuss the limitations of the available data. It is beyond the scope of this article to collect entirely new data on evaluation cultures. Instead, we merely add one country and put existing data to work in answering the following question: Do variations in evaluation culture across countries also correlate with variations in good governance in a way evaluators would like to see?

Our hypothesis is that in a sample of countries, those who have a high score on the institutionalization of evaluation also tend to have a relatively high score on good governance. Even if several threats to a direct causal interpretation of this link should be taken into account (including reciprocal causality), we believe that most evaluators would prefer to see a positive correlation between evaluation and good governance regardless of control for other factors. As far as we know, the empirical correlation between these two phenomena has not been established before. It is an important thing to do, however, in order to gauge the validity of one of the most important narratives in the field of evaluation.

However, even if empirically established, a general correlation between evaluation culture and transparency may not have the same relevance and meaning across countries. In fact, one of the classical connotations of culture has to do with variations in meaning across national contexts. Many factors influence the reception and use of evaluation in a given country. Evidently, we hypothesize that one important factor is whether a given country is already characterized by a particularly low or high level of transparency and good governance. Where transparency is low, evaluation is likely to be in demand because the room for improvement is great. In countries where transparency is high, the contribution of additional evaluation may be perceived as questionable.

This argument helps us select two countries for further study, South Africa and Denmark. We intend to show that the different positions of two countries in our plot of evaluation culture and transparency to a large extent help us make sense of evaluation debates in these countries and help explain differences in how hospitable these countries are for (more) evaluation. In other words, we argue that the meaningfulness and the relevance of the same general data may differ depending on national context. Furthermore, the location of a country in an evaluation culture/transparency graph might be a defining element in that context.

Our contribution in this article is twofold:

1. We articulate the story about how evaluation culture contributes to accountability and good governance as a hypothesis and test it empirically.
2. We demonstrate that the relevance and meaningfulness of the general narrative differ across countries. We do that by looking at evaluation histories in two countries with starkly different scores on transparency.

Both of these contributions are relevant for evaluation and evaluators. It is important to carry out research on evaluation so that the empirical basis for general narratives about evaluation can be gauged. It is also important in evaluation practice to pay attention to national contexts and other specificities which may make general narratives more or less relevant in a given evaluation situation.

We proceed in the following way: First, we explain our understanding of narratives, and specifically the relation between narratives and data in a policy context. Second, we discuss evaluation culture and how it can be measured. Third, we discuss transparency as an aspect of good governance and how it can be measured. Fourth, we show the correlation between these two variables in a sample consisting of 20 countries and discuss it in light of our hypothesis. Fifth, we provide small country case studies (South Africa and Denmark) and make sense of these case studies in dialogue with our general analysis of the link between evaluation culture and transparency. Finally, we conclude and draw perspectives.

Narratives

According to Meyer et al. (1994), modernization of the world takes place through rationalized myths about progress, justice, and rationality. They understand "rationality" here in a Weberian sense. Weber described how modern life becomes structured through impersonal rules grounded in overarching collective purposes (Meyer et al., 1994: 20). As such, modernizing narratives about progress, justice, and rationality are powerful principles of social structuration. By "myths" we mean something which is neither totally true nor untrue, but something which helps restructure reality in particular ways.

To be operational in practice, rationalized myths need to be told and retold as stories. They are general, abstract, and condensed versions of reality. They are what Fischer (2003) calls "flat" stories. Their abstraction allows them to fly across time and place in a way which is disembodied from the lived reality of concrete countries, concrete contexts, and concrete human beings.

When "the rubber hits the road," there is more or less friction. Some rationalized myths fit well with data; some rationalized myths lose credibility when confronted with "inconvenient" empirical realities (Fischer, 2003: 177).

We also live in times when some narratives are very stubborn and refuse to give up; instead people invent "alternative facts" which confirm the narratives. This does not mean that the work to establish facts is in vain. It just means that the relation between ideas and facts is characterized by much elasticity. This elasticity is caused by differences not only in ideological beliefs but also in the availability of data, the trustworthiness of data, the boundaries of time and space in which data are said to be valid, and the choice of macro/micro perspective. Furthermore, even if a narrative fits "generally" with some data in an empirical field defined in time and space, it does not mean that it fits with all concrete experiences made in that field. For these reasons, a given rationalized myth can be subject to much resistance, reinterpretation, and so-called "implementation problems" when it meets practice. People will continue to discuss how well the rationalized myth fits with the local context. In turn, the "context" can also be interpreted in different ways (Dahler-Larsen and Schwandt, 2006) so that it resonates more or less with the requirements of the general rationalized myth.

To live in a democracy is to be thrown into a situation where there are multiple competing interpretations of reality (Stone, 2012). These interpretations are narratives which struggle (more or less) with data. In a deliberative democracy, citizens "help" each other make sense of various

data in the light of competing narratives, and sometimes, consensus coalesces around particular versions of the world or particular policy frames. In other situations, competing or conflicting frames continue to exist. We also know from the sociology of knowledge that data are in practice not only evaluated on a truth criterion but perceived relevance is of tantamount importance, too. Among the many stories circulating in democratic societies are stories about evaluation.

Evaluation

Evaluation taps into modernization. Evaluation is a vehicle for modernization. Evaluation hinges on values such as professionalism and transparency (Donaldson and Lipsey, 2006). The classical purposes of evaluation, such as enhancing accountability, learning, enlightenment (or knowledge-building), and information of the public (Chelimsky, 2006), fit nicely into a broader modernization agenda. These purposes are narratives about the role of evaluation in society. One of the most powerful of these stories is the one about how evaluation enhances transparency and good governance.

If evaluation works according to one of its main official purposes, it enhances transparency and thereby promotes accountability. With transparency and accountability, corruption can be reduced and government efficiency can be enhanced. Other functions of evaluation, such as the learning and enlightenment functions, may enhance government efficiency and good service delivery over time and at different organizational levels. Therefore, the provision of evaluative data is important also for evaluators who might emphasize inclusion, dialogue, and deliberation rather than the accountability function of evaluation (Benjamin, 2015; Fraser and Rogers, 2015). In turn, these kinds of involvement of citizens may also contribute to citizens' insights into the operations and outcomes of government. All in all, there is sufficient reason to believe in a very general narrative about a link between evaluation culture and transparency.

We expect this narrative also to have an elastic relation to data and experiences. There may be tensions between upward and downward accountability in various parts of the managerial chain of control as well as tensions between accountability and learning (Benjamin, 2015).

Nevertheless, in the broader picture, most evaluators would prefer a clear and strong positive relation between evaluation culture and transparency, with or without control for other variables. Here, evaluators should remember to maintain critical thinking, and a balanced critique of positive as well as negative findings, even if it is their own narrative which is now under evaluation. We expect some elasticity here, too. Depending on how our correlation looks, evaluators will discuss the reasonableness of our measurements, the additional factors we should have controlled for, and, what we clearly admit, the possible variations in meaning and relevance of our findings in various national contexts. We now turn to measurement and data regarding evaluation culture and transparency.

Evaluation culture

Evaluation culture is a broad concept. There are at least two broad ways in which the concept is used. The first has to do with variation in meaning and sense-making. Evaluation culture describes evaluative mindsets and how people make evaluation fit into local beliefs and traditions. This dimension would also allow evaluation culture to take different forms depending on fitness for purposes defined locally. In this sense, evaluation cultures can be compared, but not ranked.

The second meaning has to do with the comparative degree of institutionalization of evaluation. Here, evaluation maturity may be the guiding metaphor. Evaluation culture thus becomes an ordinal variable so that countries can be ranked according to their degree of maturity. It is assumed that evaluation culture develops in a one-dimensional way toward serving the same higher purpose(s) wherever it occurs. This understanding of evaluation culture does not harbor adaption to variations in local purposes nor does it take into account that very different institutional configurations can all lead to good governance (Rothstein, 2012: 151). The implication is that an indicator of evaluation culture might ignore or misrepresent functional alternatives to the institutional arrangements imagined by those who defined "mature" evaluation culture in a general way. Despite these shortcomings, most measurements of evaluation culture subscribe to the second of these views. However, it is not easy to deliver an operationalization of evaluation culture for comparative purposes.

A first attempt to empirically map and rate evaluation cultures across the globe was done by Furubo et al. (2002) in "The International Atlas of Evaluation." Here, national evaluation experts involved in the Inteval network around Ray Rist rated nine aspects of evaluation culture in 21 countries according to a common schematic frame of reference. These aspects include, for example, whether the national audit office works actively with evaluation, whether evaluation is used systematically in parliament and in the administration, whether there is an active debate on evaluation in the country and a professional organization for evaluators. Selected experts could ascribe zero, one, or two points to each of these aspects, so the maximum score would be 18 points.

The approach leaves much to discuss. Where do the nine universal aspects come from? Can they be trusted as universal keys to evaluation culture, or do they in fact reflect the minds of individual editors who extrapolate their experiences from particular countries? Is every aspect of evaluation culture equally important in each country? Are the national experts truly independent or in fact evaluating their own efforts to foreground evaluation? Furthermore, the data set is limited. Only 21 countries were included in the analysis.

Nevertheless, there is no perfect way to measure evaluation culture. The International Atlas of Evaluation was trail-blazing. A follow-up including 19 countries was made in 2011 by Jacob et al. (2015). The methodology remained basically the same. The authors continue to base their scores on expert assessments. Although this method gives room for insight, expert views are subject to personal bias and may suffer from lack of reliability (Fitzpatrick et al., 2004: 125). Jacob et al. (2015: 9) acknowledge these strengths and weaknesses of expert views. To compensate for some of these weaknesses, they involve several experts per country.

The score of some prominent countries has fallen, for example, that of United States from 18 to 16 points. One wonders whether the evaluation culture of this forerunner is on retreat or whether the new figure is a result of the fact that there is no longer an American member of the editorial team. At this point, we deliberately stop our critical questions, because it is easy to criticize work that carries out difficult tasks, and because we know of no better data on evaluation culture which can be used as an alternative. We respect the work of Jacob, Speer, and Furubo as the best one available.

We add South Africa to the initial set of 19 countries, following a procedure which uses the same nine indicators of evaluation culture to ensure comparability. We also relied on multiple expert views. In the first round, we derived a score representing the current degree of institutionalization and maturity of evaluation culture in South Africa using five groups with at least one senior expert in each. Their average score served as a benchmark or point of reference.

We then asked two evaluation experts to recalibrate their initial scores to reflect the state of evaluation culture in 2011, and triangulated the resulting scores against peer-reviewed literature and in light of the known benchmark score for 2017. Based on the historical trajectory of the evaluation in South Africa, we expected our recalibrated and triangulated score to be several points below our benchmark score. See Appendix 1 for details. Given the profile of participants involved in these two rounds, and given the described triangulations and adjustments, we are confident that the method used to score the evaluation culture in South Africa is at least as valid as the methods used in previous studies of evaluation culture. We also believe it is robust in the sense that it is difficult to imagine an alternative method with the power to systematically change the scores on enough indicators to substantially alter the relative position of South Africa in our evaluation culture-transparency graph.

Transparency and good governance

Good governance includes multiple dimensions such as the rule of law, government efficiency, democracy, transparency, accountability, absence of corruption, and absence of clientelism (Rothstein, 2015). The notion of "good" refers to normative dimensions which are sometimes under-theorized (Rothstein, 2015). Some normative theories of good governance include citizen well-being and social equality (Rothstein, 2015).

Transparency International publishes every year the CPI (Corruption Perception Index) as a measure of good governance. Corruption is addressed indirectly. By definition, actual corruption is difficult to document. When perpetrators are found, exposed, and perhaps prosecuted, at least some parts of the legal-political system actually work. The least faulty method to establish the level of corruption in a country is therefore to rely not on incidences of corruption but on how people with insight judge the quality of governance. Transparency International draws on views from a number of experts and organizations (such as the World Bank).

Although not all aspects of good governance are captured by CPI, we argue that transparency, accountability, and the ability to combat corruption are important aspects of governance. While good democratic governance is more than transparency, and transparency does more than merely prevent corruption, transparency represented in the CPI is in fact a fairly good indicator of good governance since the many structural, regulative, and normative institutional pillars which enhance accountability and prevent corruption logically overlap with sound institutional arrangements in democratic regimes.

CPI is in fact published as an indicator of good governance and accepted by many as a good proxy. CPI taps into institutional arrangements supporting good governance, which are relatively stable over time. It is also correlated with many peoples' intuitive perception of variations in credibility and transparency in governments around the world. Arguments about the contribution of evaluation to good governance can be built around the centrality of transparency as a key dimension in good governance in itself, but also as a predictor and an outcome of institutional arrangements characteristic of good governance. While all effects of evaluation upon good governance may not be within empirical reach at the present moment, it is sufficient for the argument in this article that evaluators are interested in whether evaluation culture plausibly influences the kind of good governance represented by CPI. We use CPI data from 2011 to match the year of the evaluation culture exercise cited above. For practical reasons, and since we empirically rely on the CPI, we will use the terms "transparency" and

"good governance" interchangeably, while what we precisely mean is "transparency as an indicator of very important although not all aspects of good governance."

Correlation between evaluation culture and CPI

It is a fundamental lesson on page one in social science methodology textbooks that correlation is not the same as causation. So, we do not jump to making causal claims.

Our primary interest is whether data lend support to stories about evaluation. Narratives do not need perfect data. A positive correlation, hopefully strong and clear, between evaluation culture and good governance is a first step to indicate support for a positive narrative about how evaluation contributes to good governance. When the two are correlated, it is possible when working in country X to point to country Y and say: Use country Y as an example. They have a more mature evaluation culture and (therefore) also better government. Even when used only as an assumption, the underlying causal story fares much better when there is a correlation that appears to back it up.

Before making a genuine and scientifically robust causal claim out of this correlation, caveats should be considered. The first caveat is that maybe the causal link between the two is reciprocal. Evaluation begets good governance, but good governance also paves the way for evaluation. Even if this may be true, evaluators will argue that active engagement in evaluation is important and justified. A reciprocal causal relation between two variables does not constitute an argument against action on one of them. (Just because variations in weight can be shown to influence variations in exercise, as well as the other way around, you would still exercise to reduce your weight).

Then, of course, there are a large number of other factors in addition to evaluation which influence good governance. These factors may confound the correlation between evaluation and transparency. However, the story about evaluation and transparency would be less straightforward in the following version: "If you control for a large number of factors in a very complicated analysis, using statistics that few people understand, there is a really a weak underlying correlation here which supports our view." On the contrary, if there is a raw, positive bivariate correlation between evaluation culture and transparency, evaluators would probably not hesitate to refer to this finding in their stories about evaluation.

Our findings are shown in Figure 1.

We abstain from excessive statistical maneuvers on this little data set. We have merely inserted a Tukey tri-split median-based line to facilitate the visual interpretation of the data.

As this line indicates, a more mature evaluation culture is generally followed by more transparency in government. At a closer look, however, this tendency is more pronounced in the left part of the graph than in the right part. It is difficult to see any positive correlation between evaluation culture and transparency once an evaluation culture score of about 14 has been passed. Countries with a very mature evaluation culture include the United States, Canada, and the United Kingdom which have lower CPI scores than say Denmark and Sweden which have a bit more moderate evaluation culture scores. Maybe it is difficult to capture in our rough measures the finer differences in scores on the two scales (the right side of the graph is more crowded than the left side of the graph). Maybe our data set is too small. At the very least, however, our data are not inconsistent with the idea that perhaps there is a weakening or vanishing link between evaluation culture and good governance once a certain threshold has been passed.

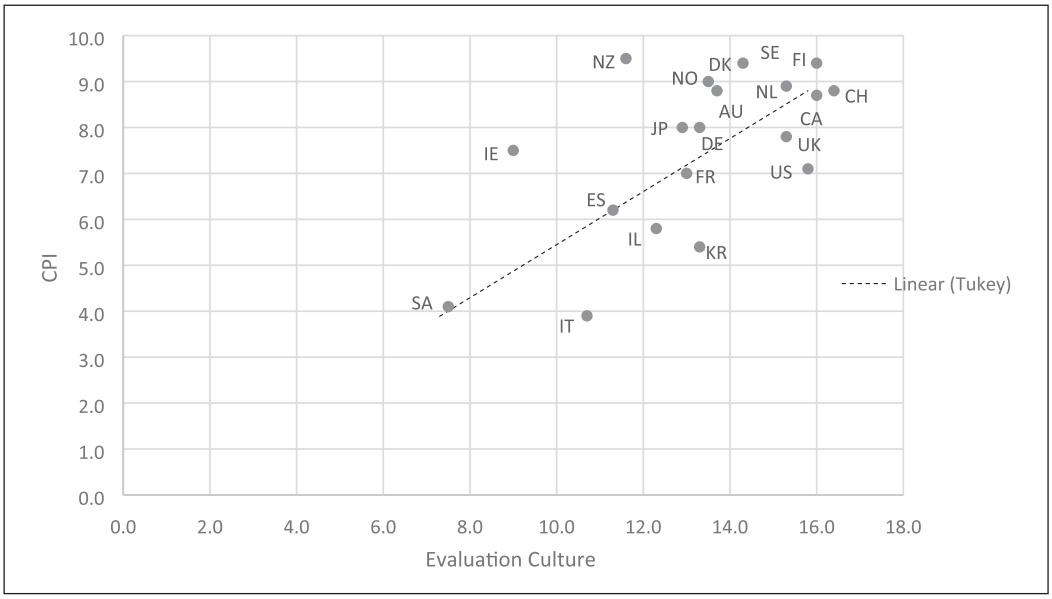

Figure 1. Evaluation culture and transparency.
Source. Jacob et al. (2015); Transparency International (2011).
Note: The dotted line represents Tukey's tri-split line (Tukey, 1977). AU: Australia; CA: Canada; CH: Switzerland; DE: Germany; DK: Denmark; ES: Spain; FI: Finland; FR: France; IL: Israel; IE: Ireland; IT: Italy; JP: Japan; KR: South Korea; NL: Netherlands; NO: Norway; NZ: New Zealand; SA: South Africa; SE: Sweden; UK: United Kingdom; US: United States.

In order to explore more contextualized meanings of evaluation culture and transparency, we now visit two countries at either end of the CPI score.

South Africa

South Africa's evaluation culture score of 7.5 is relatively low. What follows is a brief account of South Africa's evaluation landscape prior to and immediately after 2011, in an attempt to contextualize this score. We appreciate that the country's monitoring and evaluation[2] (M&E) capacity, systems, and practices have grown leaps and bounds post 2011. Our intention is not to dismiss the major developments captured and celebrated in Abrahams (2015), Goldman et al. (2018), or Phillips et al. (2014). The timing of these changes is critical; however, since our comparative exercise is driven (if not constrained) by Jacob, Speer, and Furubo's 2011 data set, our main focus is the time up to and around 2011.

Although M&E entered the South African landscape through the donor community in the early 1990s, it only gained ground when the public sector launched various mechanisms with accompanying legislative mandates to institutionalize this practice (Abrahams, 2015). Prior to 1994, public sector monitoring and reporting practices were in place to generate information for control purposes. Post-1994, initial efforts to draw on this information for decision-making purposes were undertaken. This attempt resonated with the democratic government's agenda to transform public sector operations. The transformation process unfolded in three distinct phases, namely, the rationalization and policy development phase (1994–1999), the

modernization and implementation phase (1999–2004), and the accelerated implementation phase (2004–current) (Madzivhandila, 2010). The most concrete efforts to institutionalize M&E practice in the executive arm of the state occurred in the third phase, following the conceptualization of the Government-wide Monitoring and Evaluation System (GWM&ES) in 2005. The GWM&ES was envisaged as a "system of systems" to streamline and synchronize M&E activities across the governmental departments and strengthen linkages between the Presidency, the National Treasury, and the National Statistics Agency (Abrahams, 2015; Beney et al., 2015; Phillips et al., 2014).

The GWM&ES Policy Framework was released in 2007, and while there were emerging pockets of practice in isolated sectors such as the Department of Social Development (DSD) and the Public Service Commission (constitutional entity reporting directly to Parliament), there was no overarching strategic evaluation mandate or evaluation system at that stage. Pressure on the 2009 incoming government (in the form of widespread service delivery protests) strengthened the ruling party's resolve to prioritize the M&E agenda, with a view to improve the performance and transparency of public sector service delivery systems (Goldman et al., 2015; Umlaw and Chitepo, 2015). The Department of Performance (later renamed Planning) Monitoring and Evaluation (DPME) was established within the Presidency in 2010. The designation of the DPME as custodian of the National Evaluation System (NES) and the articulation of national evaluation priorities in the National Evaluation Plan (NEP) are major turning points in South Africa's brief history of M&E institutionalization (Beney et al., 2015; Engela and Ajam, 2010). The roll out of the NES, following Cabinet's approval of the National Evaluation Policy Framework and the creation of the Evaluation and Research Unit (ERU) within the DPME in 2011, is captured in Goldman et al. (2015) and Goldman et al. (2018). The first NEP evaluations and initial deliberations about evaluation standards and competencies to support the implementation of the NES only started in 2012.

A number of other noteworthy developments unfolded in 2012, including (a) the publication of the draft of the South African Standards for Evaluation in Government, (b) the formalization of the DPME–SAMEA partnership (the South African Monitoring and Evaluation Association is a voluntary organization for professional evaluators established in 2005), and (c) a first systematic attempt to capture how M&E systems are faring in 96 national and provincial departments by the DPME in collaboration with the Deutsche Gesellschaft für Internationale Zusammenarbeit (Beney et al., 2015; Leslie et al., 2015; Umlaw and Chitepo, 2015). The findings of the 2012 situational analysis speak to a number of deficiencies in the institutional environment that compromised the implementation of M&E functions. These findings resonate with some of the earlier observations made by Cloete (2009), Engela and Ajam (2010), and Madzivhandila (2010), such as (a) few provincial departments with M&E units fully capacitated in terms budget, staff, and systems; (b) the existence of a compliance versus learning culture within the public sector, with a focus on monitoring outputs at operational level; and (c) the lack of coordination among core M&E stakeholders and "turf battles" between departments.

While the public sector is still dealing with many of these challenges (Goldman et al., 2018), we would like to reiterate that substantial progress has been made on the M&E front since 2012, and especially in the last 4 years. For example, South Africa is working toward institutionalizing the use of evaluation results in the national budget process. First, attempts were made during 2017/2018 and 2018/2019 budget cycles. The last few years have also witnessed an upsurge in the number of dedicated evaluation courses offered by higher education institutions and the emergence of a vibrant culture of evaluation research in South Africa.

The change in our evaluation culture scores from 7.5 in 2011 to 12.2 in 2017 reflects the fact that we are observing M&E in its formative years in South Africa, as a profession, industry, or governance tool. After all, the foundation for a NES was only established in 2011.

The country's positioning on Figure 1 is no coincidence and certainly does not come as a surprise, given its low CPI score of 4.1 in 2011. South Africa's 1994 transition from a 40-year-old apartheid regime to a constitutional democracy remains one of the most striking political transitions in history. Despite having one of the most progressive constitutions in the world and being at the pinnacle of transition toward more accountability and transparency, the country is fraught with corruption. Political corruption is firmly entrenched in the day-to-day reality of the country. Without discussing specific corruption cases or "scandals" that are currently making headlines and why they occur in South Africa, a crude lay man's observation would be as follows: The country's current political landscape is defined by the ongoing public controversy tied to flagrant instances of corruption.

Given the country's current positioning on Figure 1, should South Africa learn from countries with a higher evaluation culture score, with a view to improve its CPI score through accountability and transparency? There is a high expectation that evaluation might consolidate democratic governance and accountability in South Africa – the recent work of Cloete (2017), Goldman (2017), and Fraser and Rogers (2017) are, after all, built on this very thesis. Before we jump to a rash conclusion, let us look at the case of Denmark.

Denmark

Denmark has a moderate score of 14.3 in evaluation culture, but a very high CPI score of 9.4. In this light, let us give a brief account of evaluation in Denmark.

Denmark was not among the first countries to adopt evaluation (Furubo et al., 2002), but once it happened from the 1980s onward, evaluation became institutionalized in many ways. Legislation requiring the mandatory publication of evaluative data for schools was passed, particularly following an Organisation for Economic Co-operation and Development (OECD) report in 2004, which recommended an "evaluation culture" in Denmark. In the following years, schools were required to produce "quality reports" for management use, to publish average exam grades for students, and to use tests systematically (although without publication). In upper secondary schools, hospitals, and other sectors, publication of evaluative data are now in place, as well as a number of evaluative practices, such as auditing, accreditation, and monitoring (Dahler-Larsen and Hansen, forthcoming).

The National Audit Office conducts evaluations and advices public organizations in general about evaluation. There are a number of agencies and institutes which carry out evaluation and provide input for public management and policy at various levels. Evaluation has been massaged into local governments which are responsible for a large part of the welfare services in Denmark.

Denmark also has a thriving professional association for evaluation that arranges annual conferences, meetings, and so on. At the same time, however, there has also been considerable, manifest, and widespread discussions about evaluation. For example, public servants who played a part in the introduction of evaluative practices wrote a newspaper article where they deeply apologized, arguing that "we did not know what we were doing" (Gjørup et al., 2007). This was only one input into an ongoing debate about whether there is too much evaluation in the public sector.

A survey among teachers found that Danish teachers were more skeptical about the value of international assessments of school quality than teachers in Sweden, Finland, England, and Scotland. Denmark was the only country in which more teachers generally found that international assessments of school quality did more harm than good (Ozga et al., 2011).

In the same vein, a survey among groups of public employees (Dahler-Larsen and Pihl-Thingvad, 2014) found that many professionals strongly felt that evaluation and performance indicators did not sufficiently grasp the essence of what their professional work was about. Many found that performance indicators and evaluation indicated a lack of trust from the rest of society. The so-called "Danish Quality Model" in hospitals was conspicuously dismantled after two rounds of accreditations. It had created much discontent among professionals, partly because of the documentation burden put upon their shoulders. Medical doctors are writing newspaper columns explaining that the current documentation pressure encourages them to focus on outgoing patients for whom doctors have a documentation responsibility at the expense of incoming patients in acute need of treatment (Christensen, 2017).

Danish evaluation researchers have contributed to a critical public debate about evaluation. They have argued that evaluation models are sometimes chosen for reasons which are not rational (Hansen, 2005), that evaluation is sometimes a ritual (Dahler-Larsen, 2012), and that it sometimes has unofficial, constitutive consequences (Dahler-Larsen, 2014). A literature review from the municipal research institute KORA points to side effects of indicators in employment services and social work (such as creaming), working against the interest of weaker clients (Møller et al., 2016).

Danish evaluation researchers who develop evaluation models (such as Hanne Krogstrup who uses focus groups in bottom-up oriented user evaluations) focus more on learning and dialogue than on accountability and transparency. This is consistent with other Scandinavian evaluation researchers (such as Anders Hanberger, Christina Segerholm, Ove Karlsson, and others). In other words, in Denmark, a fairly extensive institutionalization of evaluation has regularly been accompanied by much debate and skepticism. Although evaluation is generally on the march forward in terms of institutionalization, there are occasional setbacks, such as the dismantling of the "Danish Quality Model" in hospitals. There continues to be, among many, a widespread skepticism about evaluation.

Looking again at Denmark's location in Figure 1, such skepticism actually makes sense. Denmark already has a high CPI score, among the highest in the world. Denmark is an outlier above the general trend line, so there must be additional reasons for the high CPI score that are not due to evaluation.

Denmark had a relatively corruption-free and relatively well functioning public sector (in terms of transparency, good governance, and public trust) before evaluation was invented and introduced. Historically, the local presence of public authority figures (such as the priest and the teacher) has been part of the experience of Danes for hundreds of years, which has probably contributed to a high degree of trust in the public sector. Denmark has a decentralized government, a fairly egalitarian society with low power distance, and a tradition for public involvement in government. Denmark also has a high educational level. Given the large extension of the welfare state, many citizens are in fact themselves public employees. A relatively corruption-free public sector may help recruit the most honest people to public employment, thus sustaining a positive circle (Barfort et al., 2016).

The Danish administrative culture has been characterized as "weak on principles" (Hansen and Beck Jørgensen, 2009). There has never been a revolution where all constitutional

principles were reinvented from scratch. Instead, there is a fairly pragmatic approach to public management consistent with a relatively low degree of formal structuration (Hansen and Beck Jørgensen, 2009). Major interests in society are often consulted before legislation is enacted. Informal dialogue is also a guiding principle regarding the overall control of the size of the public budget (Mouritzen, 2012).

Together, these factors help explain how a CPI can be achieved which is much higher than would be expected when looking at the general correlation between evaluation culture and CPI. Let us assume, for the sake of the argument, that Denmark would seek inspiration from nations with a higher evaluation culture score. The obvious choice here is United States, because of its strong development of evaluation as a distinct field with evaluation models, famous evaluation authors, strong consulting companies, and an effective distribution of ideas and literature to the rest of the world. However, American culture is more dominated by competition and power distance than the Danish one. Danes would presumably perceive evaluation paraphernalia such as scores and ranking as somewhat "foreign" to their tradition (which they in fact often do). Even more importantly, in return for becoming more like the United States, Denmark would achieve a more "mature" evaluation culture, but would at the same time lose its unique high CPI score. That would be a very unfortunate trade-off if evaluation culture is a means and CPI an end.

A moderate level of evaluation enthusiasm is thus reasonable and justified: There is already in Denmark a high level of CPI; most other countries with a more mature evaluation culture have lower CPI scores. In this regard, evaluation skepticism finds a basis in data. Given the high level of trust already, it is quite logical that evaluation is portrayed by many as an indication of "lack of trust."

Discussion and perspectives

Evaluation culture cannot be measured perfectly. We have used the best measure we could get our hands on. We have also used the best measure of good governance we could get, which is CPI. It is stable and robust and often referred to. Among our sample of 20 countries, we found CPI scores to be positively correlated with evaluation culture. This correlation lends support to a general narrative about how evaluation contributes to transparency as an indicator of good governance.

Correlation is not causality. Before making causal conclusions, reciprocal causality should be considered, and additional variables should be controlled for. A perfect causal analysis is not likely to be available soon. In the meantime, it makes sense to understand the correlation between evaluation culture and transparency.

Many other factors contribute to good governance. In a similar vein, enhancing transparency is not the only purpose of evaluation. Not all justification for evaluation hinges on its contribution to transparency. For example, contributions to reflexivity and learning also count. Nevertheless, the contribution of evaluation to transparency has been and continues to be an important narrative in evaluation.

We have found the general correlation between evaluation culture and transparency to be positive. We wish to discuss two substantial reservations. One has to do with outliers. A further look at outliers can stimulate research questions. Why are some countries way above or below the trend line? More specifically, for example, one question we heard during our data collection workshop is as follows: Why do New Zealand and South Africa have so different CPI scores? They both have a history of British colonialism, and both have indigenous populations.

Our second reservation is as follows: After a certain score in evaluation culture is achieved, a higher score no longer seems to be associated with increasing CPI. Could it be that evaluation culture is instrumental in bringing about good governance, but the effect withers away once a certain level has been achieved? If this were the case, the effect would be similar to what is known in economics as "decreasing returns to scale." This phenomenon describes that after a point, more inputs of a production factor does not lead to proportionally the same increase in production. The effect flattens out. At a point, it may even become negative. Could it be that in some of the countries with a high evaluation culture score, the auditability of public organizations has become a purpose in itself (Power, 1997) at the same time as side effects of performance measurement and evaluation are beginning to flourish, such as unintended consequences, performance paradoxes, and constitutive effects? (Dahler-Larsen, 2014; De Bruijn, 2002; Van Thiel and Leeuw, 2002).

This observation might very well be restricted to the aspect of evaluation which has to do with accountability and transparency. In political cultures where there is already relatively high trust in government, the introduction of additional evaluation may not enhance further trust, but perhaps lead people to believe that their trust is ill-founded. It might well be that after a certain level of accountability has been reached, there is little to gain here, and more emphasis should be put on learning and deliberation.

We do not have data to determine whether "decreasing returns to scale" is also true for evaluation purposes such as learning, knowledge-building, and information to the public.

However, such a shift in emphasis does not remove all demands for accountability, nor the contribution of evaluation hereto. An advanced evaluation culture would thus be able to incorporate reflexivity, taking into account the positive and negative consequences of evaluation itself with its multiple functions. This would be a more challenging and complex endeavor than merely to install more evaluation culture or a more "mature" evaluation culture. As a result, there would be two different versions – at either end of our graph – of the narrative of how evaluation and transparency could be linked. At the very least, the correlation between evaluation culture and transparency among the 20 countries studied should not be taken as a proof of the idea that further enhancement of evaluation culture at all levels of evaluation culture will automatically guarantee further improvements of transparency.

Our data set is small. Further research and a larger data set with countries having various scores on evaluation culture, modernization, and good governance will show more light on the general contribution of evaluation culture, whether it withers away at high levels of evaluation culture and whether additional factors can systematically explain outliers.

Conclusion and implications

We have confirmed, in broad strokes, our hypothesis about a positive correlation between evaluation culture and good governance. In a very general comparative perspective, this correlation can be used as an argument in favor of evaluation. However, we have found considerable reasons to modify an unqualified belief in this correlation. After a certain score in evaluation culture is achieved, a higher score no longer seems to be associated with increasing CPI. Although a demand for more evaluation in order to enhance transparency in less "mature" evaluation cultures may be well justified, evaluators in highly "mature" evaluation cultures should perhaps not ask for more evaluation, but for a more fine-tuned dose of evaluation, and for forms of evaluation qualitatively fit for other purposes than transparency.

The general, "flat" story about how evaluation contributes to transparency and good governance deserves to be told and will presumably also be told in the future. However, the narrative which springs from the correlation between evaluation culture and transparency does not resonate in a uniform way in all contexts. Evaluators should not only base their promotion of evaluation upon general ideas and correlations. The story about how evaluation contributes to transparency should be interpreted, modified, and refined in context-specific ways.

By comparing two countries located in two different parts of our evaluation culture/transparency graph, we have shown that the meaningfulness of an overall narrative about evaluation depends starkly on context. Whether one wants to analyze a given country as a context for evaluation, or whether one is seeking to promote evaluation culture in that country, a good starting point is an understanding of how the country is located in an evaluation culture/transparency graph.

Acknowledgements

The authors wish to thank University of Cape Town and the participants in the data-collection sessions, which took place there.

Declaration of conflicting interests

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

Funding

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

Notes

1. The data points were first divided into three segments on the X axis: early, middle, and late. A trend line was then drawn to connect the intersects of the median X and Y values for the early and late thirds data segments.
2. In South Africa, the terms monitoring and evaluation are often used together or interchangeably.

ORCID iD


Peter Dahler-Larsen https://orcid.org/0000-0002-3854-8587

References

Abrahams MA (2015) A review of the growth of monitoring and evaluation in South Africa: Monitoring and evaluation as a profession, an industry and a governance tool. *African Evaluation Journal* 3(1): 1–8.

Barfort S, Harmon N, Hjorth F, et al. (2016) Sustaining honesty in public service: The role of selection. In: Midwest Political Association Meeting annual conference, Chicago, IL, 16–19 April 2015.

Beney T, Mathe J, Ntakumba S, et al. (2015) A reflection on the partnership between government and South African Monitoring and Evaluation Association. *African Evaluation Journal* 3(1): 164.

Benjamin L (2015) How can evaluation strengthen democracy? In: Podems D (ed.) *Democratic Evaluation and Democracy: Exploring the Reality*. Charlotte, NC: Information Age Publishing, 225–4.

Chelimsky E (2006) The purposes of evaluation in a democratic society. In: Shaw I, Greene JC and Mark MM (eds) *The SAGE Handbook of Evaluation*. London: SAGE, 33–55.

Christensen TE (2017) Lægerne truer med at arbejde efter reglerne [The doctors are threatening to work in accordance with the rules]. *Politiken*, 15 October. Available at: http://jyllands-posten.dk/debat/kronik/ECE9951765/laegerne-truer-med-at-arbejde-efter-reglerne/ (accessed 27 October 2017)

Cloete F (2009) Evidence-based policy analysis in South Africa: Critical assessment of the emerging government-wide monitoring and evaluation system. *Journal of Public Administration* 44(2): 293–311.

Cloete F (2017) Evaluation and democratic governance: The public management perspective. In: Podems D (ed.) *Democratic Evaluation and Democracy: Exploring the Reality*. Charlotte, NC: Information Age Publishing, 81–104.

Dahler-Larsen P (2012) *The Evaluation Society*. Palo Alto, CA: Stanford University Press.

Dahler-Larsen P (2014) Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review* 16(7): 969–86.

Dahler-Larsen P and Pihl-Thingvad S (2014) *Resultatmålinger Og Stress* [Performance Measurement and Stress]. Odense: Syddansk Universitetsforlag.

Dahler-Larsen P and Schwandt TA (2006) When evaluation meets the "rough ground" in communities. *Evaluation* 12(4): 496–505.

Dahler-Larsen P and Hansen HF (forthcoming) Evaluation in Denmark. In: Stockmann R and Meyer W (eds) *The institutionalization of Evaluation in Europe*. London: Palgrave Macmillan.

De Bruijn H (2002) *Managing Performance in the Public Sector*. London: Routledge.

Donaldson SI and Lipsey MW (2006) Roles for theory in contemporary evaluation practic: Developing practical knowledge. In: Shaw I, Greene JC and Mark MM (eds) *The Handbook of Evaluation: Policies, Programs, and Practices*. London: SAGE, 56–75.

Engela R and Ajam T (2010) *Implementing a Government-wide Monitoring and Evaluation System in South Africa*. Washington, DC: The World Bank.

Fischer F (2003) *Reframing Public Policy, Discursive Politics and Deliberative Practices*. Oxford: Oxford University Press.

Fitzpatrick JJ, Sanders B and Worthen BR (2004) *Program Evaluation*. Boston, MA: Pearson.

Fraser D and Rogers PJ (2015): Is Government´s Approach to Evaluation Deepening Democracy in South Africa? In: Podems D (ed.) *Democratic Evaluation and Democracy: Exploring the Reality*. Charlotte, NC: Information Age Publishing, 209–224.

Fraser D and Rogers PJ (2017) Is Government's approach to evaluation deepening democracy in South Africa. In: Podems D (ed.) *Democratic Evaluation and Democracy: Exploring the Reality*. Charlotte, NC: Information Age Publishing, 209–24.

Furubo J-E, Rist RC and Sandahl R (2002) *International Atlas of Evaluation*. New Brunswick, NJ: Transaction Publishers.

Gjørup J, Hjortdal H, Jensen T, et al. (2007) Tilgiv os - vi vidste ikke, hvad vi gjorde [Forgive us – we did not know what we were doing]. *Politiken*, 29 March. Available at: http://politiken.dk/debat/kroniken/art5480243/Tilgiv-os-vi-vidste-ikke-hvad-vi-gjorde (accessed 27 October 2017).

Goldman I (2017) How does government evaluation in South Africa Contribute to Democracy? In: Podems D (ed.) *Democratic Evaluation and Democracy: Exploring the Reality*. Charlotte, NC: Information Age Publishing, 105–26.

Goldman I, Byamugisha A, Gounou A, et al. (2018) The emergence of government evaluation systems in Africa: The case of Benin, Uganda and South Africa. *African Evaluation Journal* 6(1): a253.

Goldman I, Mathe JE, Jacob C, et al. (2015) Developing South Africa's national evaluation policy and system: First lessons learned. *African Evaluation Journal* 3(1): 107.

Hansen HF (2005) Choosing evaluation models. A discussion on evaluation design. *Evaluation* 11(4): 447–62.

Hansen HF and Beck Jørgensen T (2009) Den danske forvaltningsmodel og globaliseringens udfordringer [The Danish administration model and the challenges of globalization]. In: Marcussen M and Ronit K (eds) *Globaliseringens Udfordringer – Politiske Og Administrative Modeller Under Pres* [The Challenges of Globalization – Political and Adminstrative Models Under Pressure]. Copenhagen: Hans Reitzels Forlag, 36–64.

Jacob S, Speer S and Furubo J-E (2015) The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later. *Evaluation* 21(1): 6–31.
Leslie M, Moodley N, Goldman I, et al. (2015) Developing evaluation standards and assessing evaluation quality. *African Evaluation Journal* 3(1): 112.
Madzivhandila TP (2010) *A practical programme evaluation model for the Limpopo Department of Agriculture*. Doctoral Thesis, University of New England, Armidale, NSW, Australia.
Meyer JW, Boli J and Thomas GM (1994) Ontology and rationalization in the western cultural account. In: Scott WR and Meyer JW (eds) *Institutional Environments and Organizations*. Thousand Oaks, CA: SAGE, 9–27.
Møller MØ, Iversen K and Andersen VN (2016) Review af resultatbaseret styring [Review of performance-based management]. Report, KORA, Copenhagen. Available at: https://www.kora.dk/media/5186743/10946_review-af-resultatbaseret-styring.pdf (accessed 3 November 2017).
Mouritzen PE (2012) On the (blessed) deficiencies of Danish democracy. In: Blom-Hansen J, Green-Pedersen C and Skaaning S-E (eds) *Democracy, Elections and Political Parties, Essays in Honor of Jørgen Elklit*. Aarhus: Politica, 181–92.
Ozga J, Dahler-Larsen P, Segerholm C, et al. (2011) *Fabricating Quality in Education: Data and Governance in Europe*. London; New York: Routledge.
Phillips S, Goldman I, Leon B, et al. (2014) A focus on M&E of results: An example from the Presidency, South Africa. *Journal of Development Effectiveness* 6(4): 392–406.
Power M (1997) *The Audit Society*. Oxford: Oxford University Press.
Rothstein B (2012) Good Governance. In: Levi-Faur D (ed.) *The Oxford Handbook of Governance*. Oxford: Oxford University Press, 143–154.
Rothstein B (2015) Good Governance. In: Levi-Faur D (ed.) *Oxford Handbook of Governance*. Oxford: Oxford University Press, 143–54.
Stone D (2012) *Policy Paradox, the Art of Political Decision Making*. New York: W.W. Norton & Company.
Transparency International (2011) Corruption perceptions index 2011. Available at: http://www.transparency.org/cpi2011/results (accessed 28 February 2017).
Tukey JW (1977) *Exploratory Data Analysis*. Menlo Park, CA: Addison-Wesley.
Umlaw F and Chitepo N (2015) State and use of monitoring and evaluation systems in national and provincial departments. *African Evaluation Journal* 3: 134.
Van Thiel S and Leeuw FL (2002) The performance paradox in the public sector. *Public Performance and Management Review* 25(3): 267–81.
Vedung E (1997) *Public Policy and Program Evaluation*. New Brunswick, NJ: Transaction Publishers.

Peter Dahler-Larsen is Professor at the Department of Political Science, University of Copenhagen, where he leads CREME, an evaluation research center. He wrote *The Evaluation Society* (Stanford University Press 2012).

Adiilah Boodhoo is a Senior Lecturer in Organizational Psychology at the University of Cape Town. Her research focuses on comparative investigations of evaluation practice in developing and developed countries.

Appendix 1

In line with Jacob et al. (2015), we relied on subjective expert judgments to derive our initial/benchmark evaluation culture score, keeping in mind the acknowledged limits and merits of this approach. We administered a paper-and-pencil survey to five groups of participants with a broad knowledge of the South African evaluation landscape during a seminar presented by the authors at the University of Cape

Table 1. Evaluation culture in South Africa in 2011. Triangulated scores across nine idicators.

Indicator	Score
Evaluation takes place in many policy domains	1
Supply of evaluators from different disciplines	0.5
National discourse concerning evaluations	1
Professional organizations	1
Institutionalization of evaluation in Government	1
Institutionalization of evaluation in Parliament	0.5
Pluralism of institutions or evaluators	0.5
Evaluation within the Supreme Audit Institution	1.5
Proportion of impact and outcome evaluations in proportion to output and process evaluation	0.5
Total score	7.5

Town in February 2017. Our sample consisted of evaluation researchers, teachers, practitioners, and students. Each group comprised at least one participant with senior evaluation expertise.

The five groups independently rated each of the nine indicators on a scale from 0 to 1, following a brief verbal explanation of the scoring protocol, consistent with the explanatory text provided by Jacob et al. (2015). The total scores across the five groups ranged between 10 and 14, with a few instances of missing data on indicator 6 (institutionalization of evaluation within Parliament) and 8 (evaluation within the Supreme Audit Institution) accounting for the discrepancy. The missing data can be explained by the possible uncertainty/confusion around what entity would constitute the Supreme Audit Institution in the context of South Africa and the absence of relevant insiders who could provide meaningful input on these two indicators, in some of the participant groups. The resulting score for 2017 was 12.2.

Since we needed data that could be meaningfully used in conjunction with Jacob, Speer, and Furubo's 2011 data set, we asked two of our participants with the highest level of expertise and involvement in the field to recalibrate their assessment of each indicator to 2011. After this stage, we estimated the correct score for 2011 to be in the range between 5 and 10.

We then triangulated the scores for each indicator against concrete milestones in the historical trajectory of evaluation in South Africa using relevant literature. We took into account, for example, that the basic foundations of a NES was established in 2011, and that it takes years for such a major reform to be firmly institutionalized. Similarly, while the South African Monitoring and Evaluation Association (SAMEA) was in existence in 2011, it was not the vibrant and influential organization that it is today. Our final triangulated score for 2011 is 7.5 (see Table 1).

There is admittedly an element of uncertainty regarding the precision of the SA evaluation score in 2011, but a marginally alternative procedure would not lead to a result that fundamentally changes the relative positioning of the country in our evaluation culture-transparency graph nor one that would compromise our central thesis. We are also confident that our triangulated score of 7.5 does not deviate substantially from the "true" score. There is only a small subset of possible scores, all adjacent to one another, that would be consistent with the literature, the recalibrated expert views for 2011, and a historical trajectory leading up to a score of 12.2 in 2017. In the absence of "objective" data, our final score of 7.5 remains the best available approximation of evaluation culture in South Africa.