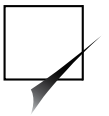
Check for updates



Article

Evaluation
2019, Vol. 25(1) 46–61
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1356389018802135
journals.sagepub.com/home/evi

Evaluation systems: How do they frame, generate and use evidence?

Žilvinas Martinaitis
Vilnius University, Lithuania; Visionary Analytics, Lithuania

Aleksandr Christenko
Visionary Analytics, Lithuania

Lina Kraučiūnienė
Visionary Analytics, Lithuania

Abstract
How does the design of evaluation systems affect the different ways of using the results of evaluations? This article offers a conceptual model that outlines three 'ideal' types of evaluation systems. It is a heuristic tool for opening up the 'black box' of evaluation systems and assessing their qualitative differences in terms of types of 'owners' of evaluations, questions asked, methods deployed, answers provided and avenues for use of evaluative knowledge. We apply the model to study the case of the Lithuanian evaluation system. In contrast to the expectations of some of the previously developed models, it does use evaluation results, and we aim to understand why the generated evidence is more often used in some areas rather than others.

Keywords
European Union, evaluation systems, ideal types, impact, Lithuania

Introduction

There is a broad consensus in the academic literature regarding why some evaluations tend to have a larger impact on the decision-making process than others. The evaluative knowledge is more likely to be used if evaluations address relevant questions, rely on robust analytical design, produce evidence in a timely manner and effectively communicate it to the relevant

Corresponding author:
Žilvinas Martinaitis, Visionary Analytics, M. Valančiaus St. 1A, Vilnius LT-03155, Lithuania.
Email: zilvinas@visionary.lt

stakeholders (Christie, 2007; Cousins and Leithwood, 1986; Leviton and Hughes, 1981; Woolcock, 2013). The characteristics of individual evaluations, however, cannot explain why some institutions, regions and countries tend to systematically produce higher quality evaluations, which are better integrated into the decision-making process.

As a result, over the past decade scholarly attention has shifted towards evaluation systems. Leeuw and Furubo (2008) distinguish evaluation systems from ad hoc initiatives according to four characteristics. First, evaluation systems rely on a shared epistemological perspective that involves an agreement among key players on objectives of evaluations and means by which they can be achieved. Second, evaluations are carried out by professional organisations specialising in this field rather than 'lone heroes'. Third, evaluations should be carried out on a permanent rather than an ad hoc basis. Finally, information obtained from evaluations should be institutionally linked to the decision-making process. This stands in contrast to the ad hoc studies that seek to satisfy one-off information needs.

Recent studies have broadened our understanding of why evaluation systems are set up (Boddington, 1993; Caffrey and Munro, 2017; Dahler-Larsen, 2012) as well as how individual systems function (e.g. Rip and van der Meulen (1995) discussed the evaluation system of the Netherlands, Lillis (2000) of New Zealand, while Barker (2007) did the same for the United Kingdom). However, there are only few papers (Højlund, 2014a; Raimondo, 2018) that examine how the characteristics of evaluation systems affect the different ways in which evaluative knowledge is used or misused.

This article seeks to contribute to the discussion in two respects. First, the article proposes a typology of 'ideal' types of evaluation systems, which builds on the synthesis of literature on knowledge management and evaluation systems. The basic argument is that the different characteristics of four key elements of the evaluation system – types of evaluation 'owners', questions, methodological designs and evidence – are mutually consistent and complementary, and have an effect on how evaluations are used.

Second, the article provides a case study of the Lithuanian evaluation system, which was set up in the run-up to European Union (EU) accession. The case study predominantly focuses on evaluation of the European structural and investment funds (ESIF). Following the proposed conceptual approach, the study seeks to understand why most evaluations in Lithuania are geared to improve implementation and management systems. This case is interesting in several respects. On one hand, the model proposed by Højlund (2014a) would suggest that evaluation systems set up due to external pressures, as was the case in Lithuania, should predominantly employ evaluations to legitimise interventions rather than improve their implementation. Hence, why are the results of evaluations are used at all in Lithuania? On the other hand, ideally evaluations should facilitate development of better policies (EVALSED, 2013; Højlund, 2014a; Pawson and Tilley, 1997a) rather than mere improvements in implementation and management. Hence, why do the impacts of evaluations in Lithuania rarely extend to policy-making?

This article follows a traditional structure. In the first section, we review the existing literature on how the characteristics of evaluation systems affect the use of the evaluative knowledge. Building on the gaps in the literature, in the second section we outline our analytical framework and the proposed 'ideal' types of evaluation systems. In sections 'Evaluation systems' and 'Methodology', we discuss the methods and findings of the case study on the Lithuanian evaluation system. The last section is where we provide a conclusion and discuss the directions for further research.

The role of external pressures, internal motivations and capacities

The literature has offered several conceptual models on how the characteristics of evaluation systems affect the use or misuse of evaluations. Building on a theory of organisation, Højlund (2014a) argues that a combination of external pressure and internal propensity to evaluate can explain the different roles evaluation findings play in the decision-making process. Strong external pressure from donors that provide financial assistance is conductive to establishing an evaluation function in the receiving countries and organisations. However, in the face of low internal propensity to carry out evaluations, the latter are likely to be decoupled from the decision-making process and the findings will be primarily used to legitimise the actions of the organisation receiving the funding from donors. If external pressure and internal willingness to evaluate are high, then organisations will avoid decoupling and some of the evaluative knowledge may be used to improve the design and/or implementation of interventions. However, Højlund (2014a) argues that such systems may also use evaluations solely for legitimisation of interventions, if they operate in a highly uncertain environment (e.g. the levels of future funding directly depend on outcomes of evaluations). On the other side of the spectrum, if external pressures are low, evaluation systems are not likely to be set up. However, if internal push towards evaluations emerges, the evaluations will be carried out on an ad hoc basis and their findings will be used to inform decision-making.

While Højlund's (2014a) approach focuses on the incentives to evaluate, Raimondo (2018) offered a model that accounts for organisational culture and resources. The basic argument is that evaluation systems can produce knowledge for policy improvement if there is appropriate internal culture (e.g. results orientation, learning culture), adequate resources and a supportive external environment. This is a tall order for an effective evaluation system. Accordingly, evaluation systems become dysfunctional when organisations in charge of evaluations face 'ambivalent signals from outside that may clash with the internal culture – and – because internal organisational processes do not incentivize evaluation use' (Raimondo, 2018: 35). Whenever this is the case, evaluations will be decoupled from the decision-making process and evaluative knowledge will be used for legitimisation purposes.

Overall, the above discussed models paint a rather pessimistic view of the capabilities of evaluation systems to produce knowledge for policy improvement. This is due to two factors. First, both models are static, that is, do not account for changes in evaluation systems, which could address tensions and dysfunctions. The evolution of evaluation systems in the EU member states provides a case in point. In line with the expectations of the Højlund's (2014a) model, evaluation practices in most member states, 'old' and 'new', have emerged mostly as a result of pressures from the EU to evaluate Structural Fund programmes (see Furubo et al., 2002; Toulemonde and Bjornkilde, 2003, among others). However, it is not the case that all such externally imposed systems are formalised, or produce knowledge that is predominantly used symbolically (i.e. to secure further funding rather than feed into discussions on policy design). In fact, Toulemonde and Bjornkilde (2003) argue that in some countries (e.g. Ireland and some regions in Italy) the opposite has happened. Externally imposed evaluation systems were internalised, and now, produce knowledge that is used beyond mere legitimisation. Furthermore, in many cases evaluation systems have subsequently spilled over from EU Structural Funds to national/regional programmes, where no external pressures to adopt evaluation systems exists. This suggests that external pressure can stimulate the emergence of evaluation systems; however, once they are set up, they evolve according to a different logic.

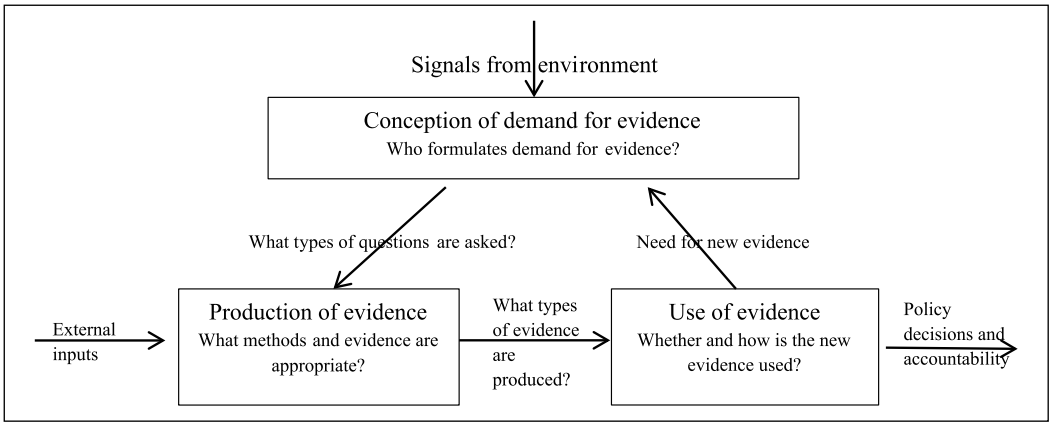


Figure 1. Evaluation cycle.
Source: own elaboration, based on organisational knowing cycle proposed by Choo (1998).

A change of evaluation systems depends on the structure and constellation of actors within organisations that are in charge of producing and channelling the evaluative knowledge. This brings us to the second point: the models discussed above do not sufficiently open-up the 'black box' of internal workings of organisations. Raimondo (2018) emphasises the importance of resources, administrative culture, and conflicting values and norms across internal units. However, to explain the diversity and evolution of evaluation systems, we need to understand whose interests and what values prevail under diverse settings and how this affects the use or misuse of evaluations.

Evaluation systems

To fill in the gaps in the literature, we propose a typology of 'ideal' types of evaluation systems. It is based on a synthesis of two streams of literature. First, we rely on the knowledge management literature to identify key elements of evaluation systems that form the backbone of the proposed typology. Second, we draw on the insights of the evaluation systems literature to characterise the elements that collectively define the 'ideal' types. Synthesis of the two streams is natural, given that evaluation systems (should) constitute an integral part of broader knowledge management systems.

The organising framework behind the typology builds on Choo (1998) and Grover and Davenport (2001), who argued that the organisational knowledge cycle consists of three elements: (1) conception of the need, (2) production and (3) use of knowledge (evidence). Figure 1 depicts the cycle adapted for the purposes of studying evaluation systems. Although all knowledge management (and evaluation) cycles follow the same phases, their characteristics differ in five respects, which constitute critical elements of the system. First, who has the right and resources to engage in knowledge brokerage? At the conception phase the knowledge brokers (i.e. evaluation 'owners') translate external pressures and internal needs into an operational knowledge production agenda (i.e. evaluation plan). Given the diversity of needs and pressures, allocation of the rights to knowledge brokerage affects the scope and direction of knowledge creation, that is, what will be studied, what questions will be posed and so on. Second, what types of questions are deemed as meaningful and legitimate? This depends on the values and

Table 1. 'Ideal' types of evaluation systems and their characteristics.

Evaluation use	No evaluation system (all types possible)	Symbolic	Instrumental	Conceptual
Owners	Evaluators	Units in charge of reporting to donors	Units responsible for policy implementation, management of funds	Policy-makers
Evaluation questions	Depends on perceived problematic issues	Descriptive/ factual questions: What happened? What are the perceptions?	Evaluative questions (e.g. relevance and efficiency)	Causal questions: Why? How?
Methodology	Any combination of methods	Monitoring data and surveys	Descriptive statistics and case studies	Advanced statistical analysis, theory-based evaluations, participatory methods
Type of evidence	Varies	Data	Information	Knowledge

Source: own elaboration.

interests of knowledge brokers (i.e. evaluation 'owners'), who frame the agenda for evidence collection. Third, what evaluation designs and evidence are deemed as acceptable at the evidence production phase? This to a large extent depends on the types of questions asked. Fourth, what type of evidence is produced? Finally, how the evidence is used or misused in the third phase (i.e. use of knowledge).

To the extent that there exists an evaluation system, the characteristics of the five elements are mutually consistent and complementary. The evaluation 'owners' have an impact on the types of questions that are perceived as legitimate, which in turn limits the scope of acceptable evaluation designs, which supports production of certain types of evidence and so on. The different combinations of these elements constitute the 'ideal' types of evaluation systems (see Table 1). Clearly, real-world systems are characterised by multiple and competing stakeholders, evaluation designs, methodologies and types of evidence produced by evaluations. Hence, the discussion below does not aim at an accurate description of a particular evaluation system. Instead, it provides a heuristic instrument that explains the differences in the inner workings of evaluation systems, and their impact on the policy cycle. Below we discuss the characteristics of each element in more detail.

The first criterion for distinguishing the 'ideal' types of evaluation systems refers to how evaluative knowledge is used. The literature on evaluation systems distinguishes between three types of evaluation use: symbolic, instrumental and conceptual (see Alkin and King, 2016; Knorr, 1977; Pelz, 1978; Rich, 1977 among others). The systems that rely on *symbolic use* carry out evaluations to provide legitimacy for organisations, policies, programmes and the like. They closely resemble Højlund's (2014a) externally enforced systems that decouple decision-making and evaluation functions.

Evaluation systems that rely on *instrumental use* are geared to collecting information that can improve management of funds and/or implementation of programmes. This type of evaluation closely corresponds to what Leeuw and Furubo (2008) characterised as an evaluation

system producing 'largely routinized information relevant for day-to-day practices and single-loop learning but of little relevance for fundamental reassessment and double-loop learning' (p. 165). According to Højlund (2014b) this evaluation use is dominant in the EU.

Systems that use evaluations *conceptually* aim to inform policy-makers regarding what works and what does not. The main objective of evaluations is to create knowledge about the cause and effect that would help to develop better programmes and policies. As the overall aim of evaluations is to change policies (e.g. see EVALSED, 2013; Højlund, 2014a; Pawson and Tilley, 1997a), such a system is closest to the 'ideal'.

Finally, we can identify situations in which evaluation systems do not exist: 'lone heroes' perform evaluations on an ad hoc basis. As there is no system in place, evaluations here can take any of the three types (conceptual, instrumental or symbolic). Examples of such arrangements include curiosity-driven or ad hoc studies.

The way in which the three 'ideal' types use the evaluative knowledge heavily depends on other elements of evaluation systems. These are: who are the main 'owners' of evaluations, what questions do they put forward for evaluations, what methodological designs are used and what evidence does the evaluations create? The next two subsections discuss each feature.

Evaluation owners and questions asked

Evaluators typically assume that the types and degree of ownership of an intervention has significant impacts on the success of its implementation (Sullivan and Stewart, 2006). We expand this idea to evaluation systems: the way evaluations will be used depends on who the owners are. Ownership here is defined as individuals or organisations that have explicit need for evaluative information, and hence they participate in defining the scope and objectives of evaluations. The degree of ownership is defined as the level of resources the 'owners' are willing to invest in the planning, implementation and follow-up of evaluations. Accordingly, the different types of owners have diverging information needs, and therefore will ask different questions the evaluation should answer.

Typically, four groups of stakeholders have a vested interesting in policy evaluation: (1) individuals or organisations that carry out the evaluation (evaluators), (2) organisations that report to donors (e.g. organisations that report on how the structural funds are being used), (3) organisations and individuals responsible for administration of funds and implementation of programmes (e.g. programme managers) and (4) policy-makers. The four groups directly correspond to the four groups of evaluation systems defined above.

When there is no evaluation system the main 'owners' are evaluators themselves as evaluations are conducted ad hoc without any explicit interest from governmental organisations. Here, evaluation questions heavily depend on what evaluators choose to research, ranging from simple descriptive needs to more complex questions of causality.

When the main 'owners' of evaluations are the organisations and units in charge of reporting to the donors, evaluations are primarily used symbolically. This is because the 'owners' aim to demonstrate that funds were spent as planned and intended objectives, given external circumstances, were attained. Accordingly, this group of owners are predominantly interested in descriptive/factual questions, such as 'What is the uptake of the programme?', 'How many outputs were created?' and 'To what extent are stakeholders satisfied?' If the main 'owners' are the organisations or units responsible for policy implementation and sound management of funds, the evaluations are likely to be used instrumentally. The stake of 'owners' in the

results of evaluation coincides with the scope of their functions: effective and efficient implementation, reaching the right target groups, improving monitoring and reporting systems and so on. Such bodies are primarily concerned with better implementation of programmes (single-loop learning) rather than implementation of better policies (double-loop learning), since the latter is beyond the scope of their jurisdiction. Hence, they are primarily interested in standard evaluative questions such as effectiveness, efficiency and timeliness.

Policy-makers (e.g. heads of policy departments, political appointees) have the interest and the mandate to review the intervention logic and to redistribute funds between the interventions. While they may face operational questions related to better implementation, they are primarily concerned with implementation of better policies. As a result, this group is interested in causal questions: why and how impacts have (not) materialised, how to design policies so that to contribute to the higher level objectives? The answers to such questions can be used conceptually, that is, to engage in double loop learning.

The 'owners' and evaluation questions discussed here also quite closely mirror the evolution of the European Commission's evaluation landscape throughout the years, as was discussed by Højlund (2015). Between 1984 and 1990, evaluations in the European Commission were carried out unsystematically, predominantly driven by evaluators' curiosity, without clearly defined guidelines. In 1995–1999 Commission produced common evaluation guidelines and made the first steps towards institutionalisation of evaluations within DGs. During 2000–2006, according to Højlund (2015), the evaluation practices in the Commission further improved with the introduction of the Better Regulation agenda (2000). Evaluations were predominantly 'owned' by operational and evaluation units, and resulted in incremental programme adjustments. The last wave of reforms started in 2007 with the view to strengthening policy learning and evidence-based policy-making. To achieve this, high-level policy-makers should increasingly share the 'ownership' of evaluations.

Types of methods and evidence

Each evaluation starts with an evaluation question. Formulation of a question largely determines the type of answer, that is, the type of evidence. Profound knowledge is necessary to answer causal questions, whereas raw data might suffice to produce answers to descriptive questions. Hence, we can assume that the methodological design of evaluations ranges from relatively simple ones that rely on a few straightforward methods (e.g. descriptive statistics, graphical analysis) to complex designs, including quasi-experimental ones, such as counterfactual analysis or participatory foresight exercises. Accordingly, the evidence delivered by evaluations can be classified into three groups: (1) data, (2) information and (3) knowledge. This classification is based on the DIKW (data-information-knowledge-wisdom) pyramid, originally proposed by Ackoff (1989). We exclude wisdom, since it deals with normative and ethical questions that are only inherent to people and thus cannot be a product of a system (Rowley, 2006).

When evaluations are primarily used symbolically, the questions are descriptive in nature and hence simple data collection and analysis methods are enough to answer them. The data collection and analysis methods most often used in such systems include monitoring data, interviews, surveys and similar. Since in-depth analysis is not conducted, evaluation reports typically provide lengthy arrays of collected data, which describes some aspects of observed cases, but prohibits any generalisation, far reaching conclusions or profound recommendations.

Hence, such evaluation systems produce data as the main type of evidence. Woolcock (2013) found that such evaluation designs are prevailing in countries where evaluation practices are in their infancy and the results of evaluations are predominantly used symbolically.

The answers to evaluation questions that are instrumental in nature mainly produce information. It is defined as organised or structured data, which has been processed so that it is made relevant, meaningful and/or valuable for a specific purpose or context (Rowley, 2006: 172). To achieve this, methodological designs rely on an analytical operationalisation of evaluation questions, clearly defined indicators and comparison of the values of cases against empirical or theoretical benchmarks so as to demonstrate the extent to which policy/programme is relevant, effective, efficient and so on.

Advanced methodological design is necessary to answer causal questions that are prevailing when evaluations are used conceptually. Most of the time evaluations seek to assess the (likely) impacts of a proposed set of alternatives. Such analysis requires counterfactual, theory-based and/or advanced statistical design. However, sometimes evaluations at this level face poorly defined problems. They are characterised by significant uncertainty over future developments, complex inter-relationships between variables and/or lack of agreement regarding interpretation of past evidence (what counts as evidence and/or the direction a policy should take). In such cases, sophisticated analytical designs fail to produce the required results and therefore participatory methods that aggregate individual views and information into shared values, meanings and knowledge are more relevant. Both designs (analytical and participatory) produce knowledge – shared practical and theoretical understanding of how interventions work, which can be generalised to other (unobserved) interventions and results from a combination of information, expert opinion, skills and experience.

Methodology

To demonstrate the feasibility and versatility of the proposed typology, we analyse the evaluation system in Lithuania. The analysis relies on the data collected through meta-analysis of evaluations, a survey of civil servants, focus group discussions and case studies. Most of the data were collected as part of an evaluation study, ‘The impact of evaluations, carried out in Lithuania between 2007 and 2014’, which was commissioned by the Ministry of Finance of Lithuania.

Meta-analysis

Meta-analysis covers 57 evaluations carried out in 2007–2017 that the EU (co)funded in Lithuania. The analysis relies on the database of evaluations and contains three groups of variables. The first group encompasses ‘demographic’ data on evaluations – title, name of the public body that commissioned the evaluation, date when the final evaluation report was submitted, type of the evaluation (operational or strategic), scope of evaluation and budget. The second group of variables lists all recommendations, organisations responsible for implementation, as well as the commitment of respective public bodies to implement the recommendation (accepted, partially accepted or not accepted). The last group of variables contains information on the type of each recommendation: (1) what does a recommendation refer to (i.e. change in policy priorities/objectives, different implementation instruments, change in administration systems, monitoring and evaluation systems or other) and (2) scale of proposed

changes (alter overall policy, introduce changes in programme, which is part of broader policy or change specific measure/instrument).

Survey of civil servants

A survey of civil servants from ministries and public organisations was the second analytical cornerstone, which provided additional information about the evaluation landscape in Lithuania. It was conducted in January 2015. In total, 171 civil servants received an invitation to participate in the survey. Around half of them completed the questionnaire. Civil servants from policy departments made up 35 per cent of all respondents, heads of policy departments at 36 per cent and civil servants responsible for the administration/management of the EU structural funds at 29 per cent. The respondent list included representatives from all Lithuania ministries and several agencies that commission evaluations.

The survey questionnaire had several levels. Level one included general questions, which were applicable to all respondents. Level two included questions that referenced specific evaluations that fall within the scope of work of the respective ministry or department. Several example questions of the survey include: whether respondents are aware of evaluations relevant to their ministry and/or department, whether they participated in the design of evaluation and discussions on its outcomes and the status of specific recommendations. In total, 28 evaluations and around 110 recommendations were used in the questionnaires. Evaluations were selected so that they cover all relevant ministries, that is, ministries that coordinated EU structural fund interventions during the analysed period.

Focus groups

The focus group discussions, organised with the members of the Evaluation Coordination Group, which is predominantly comprised from civil servants responsible for coordinating evaluations, provided a more in-depth overview of the situation in Lithuania. The main goal of the focus group discussions was to assess what types of evidence evaluations deliver.

Case studies

The case studies provided in-depth information that complemented the results of the meta-analysis, survey and focus group discussions. In total, three case studies were carried out. They focused on good evaluation practices in Lithuania. Each case study was based on interviews with civil servants, the analysis of evaluation reports and other sources.

Evaluation system in Lithuania

Evaluations in Lithuania emerged at the turn of the century in response to the formal requirements from the EU to assess pre-accession financial assistance. The first evaluations were commissioned by the EU Commission and were carried out by the EU evaluators with the assistance of local experts. Institutionalisation of evaluation function started in 2003 with the establishment of a dedicated unit in the Ministry of Finance (MoF; The Ministry of Finance of Lithuania et al., 2013). The evaluation system almost exclusively focuses on evaluations of interventions funded by ESIF. In total, 72 external evaluations of ESIF interventions were

conducted in Lithuania in 2004–2015. The MoF commissioned nearly half of all evaluations, while the remaining ones were carried out on behalf of the other 23 public bodies. The total budget of evaluations commissioned in 2004–2015 was around €4.22 million.

The MoF has the overall responsibility for coordination of evaluations. In this respect it has four main functions. First, the MoF commissions inter-sectoral evaluations, which span beyond the scope of competence of the individual line ministry or agency. Second, the MoF coordinates follow-ups to the evaluations. This includes reporting to the Commission and monitoring of the implementation of recommendations contained within evaluations. Third, the MoF chairs the Evaluation Coordination Group, which comprises representatives of relevant ministries and agencies and is responsible for the preparation and review of annual evaluation plans. Finally, the MoF carries out evaluation capacity building activities that include training civil servants, evaluation workshops and conferences, development of evaluation standards, guidelines and the like.

Line ministries that are in charge of implementing interventions funded by the ESIF constitute an integral part of the national evaluation system. They are tasked with planning and commissioning evaluations within their scope of competence as well as the implementation of recommendations that they deem as acceptable. Typically, line ministries have one or several civil servants in charge of managing the evaluations, who reside within the departments responsible for ESIF.

The sub-sections below provide a more in-depth analysis of the Lithuanian evaluation system following the proposed heuristic model of evaluation systems. Overall, the system comes close to the 'ideal' type that encourages instrumental use of evaluations. The units in charge of policy implementation and management of the ESIF are the main 'owners', they focus on standard evaluative questions (e.g. relevance, effectiveness) and the evaluations mostly produce information that is used for improving implementation.

Ownership

Most evaluations have formal or informal steering groups, which plan evaluations and draft tender specifications, provide guidance to external evaluators and follow-up with dissemination of evaluative knowledge as well as implementation of recommendations. Analysis of the composition of the steering groups provides insights into who are the main 'owners' of evaluations. Available data suggests that civil servants responsible for the management and administration of the ESIF and heads of policy departments constitute the largest groups (see Figure 2).

The results of the survey that sought to assess awareness of the results of evaluations indicate the same trend. Almost all officials employed in the departments dealing with the ESIF (92%) knew of at least one evaluation, while two-thirds of the heads of policy departments (76%) and less than half of civil servants at policy departments (46%) were aware of the findings of at least one evaluation (Figure 3). This suggests that civil servants in the units responsible for management of funds as well as policy-makers constitute the most important groups of 'owners' of evaluations.

Evaluation questions

Meta-analysis of evaluations assessed what types of questions evaluations sought to answer. It suggests that standard evaluative questions (on relevance, effectiveness, efficiency, etc.)

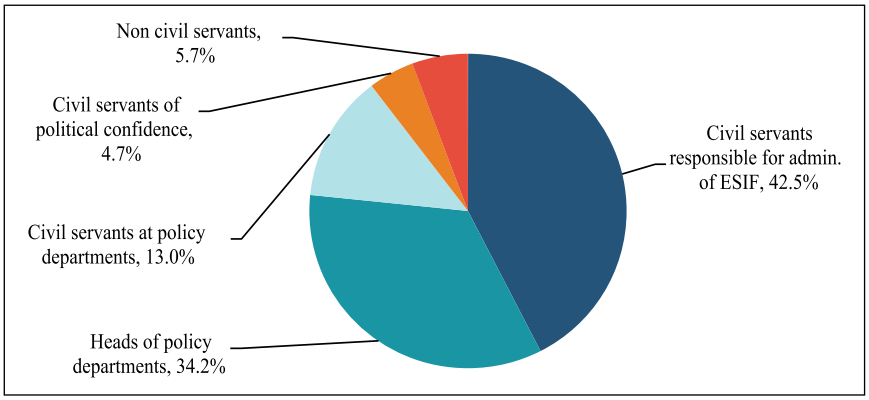


Figure 2. Composition of evaluation steering groups in Lithuania.
Figure provides data on composition of steering groups in 27 evaluations carried out in 2007–2014. Information on steering groups of other evaluations was not available.

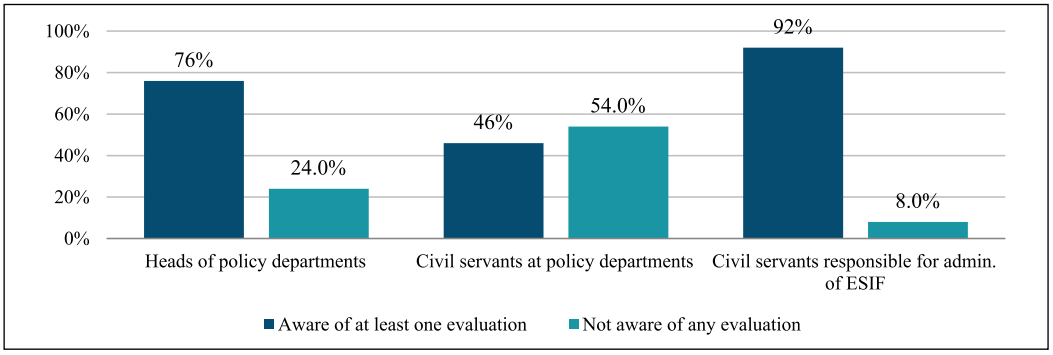


Figure 3. Officials' awareness of relevant evaluations.
N = 83 civil servants.

were dominant in evaluations carried out in Lithuania in 2007–2014 (see Figure 4). One could argue that this is not surprising, given that the EU and national evaluation guidelines emphasised these types of questions. However, the guidelines developed for the 2007–2013 programming period also sought to promote impact assessments that rely on causal questions. Our conceptual model would suggest that standard evaluative questions dominated, because they were of interest to the most populous group of evaluation 'owners' – civil servants in the units responsible for management of funds, who are predominantly interested in improving implementation of interventions. Hypothetically, more active involvement of policy-makers in planning of evaluations could have resulted in a larger share of causal questions that shed light on impacts of interventions and facilitate improvements in the overall policy design.

Methods

Results of the meta-analysis of 57 evaluations suggest that descriptive statistics, interviews, surveys and expert-panel discussions were among the most frequently used methods (see

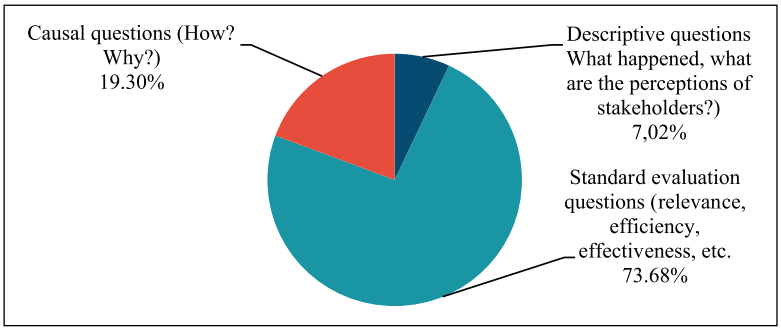


Figure 4. Type of evaluation questions.
N = 57 evaluations carried out in 2007–2014.

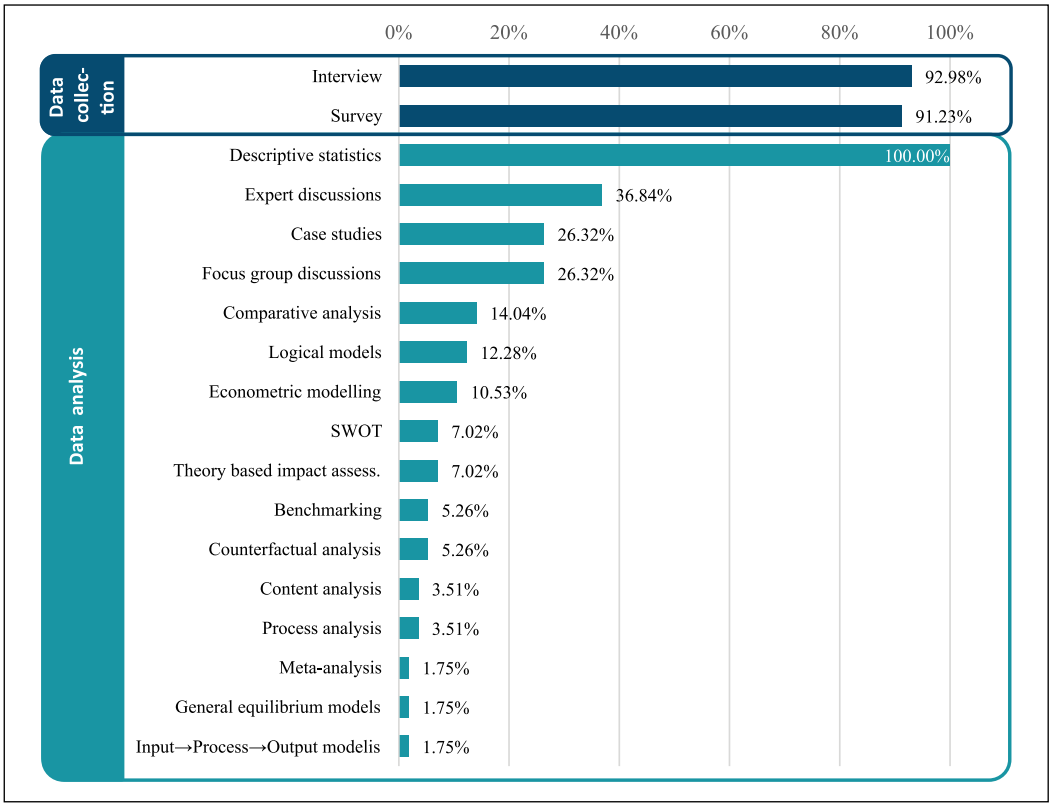


Figure 5. Methods used in evaluations.
N = 57 evaluations carried out in 2007–2014.
The methods mentioned in the bar plot were all extracted from evaluations (not our classification).

Figure 5). A small number of evaluations also used complex methods, such as econometric modelling, counterfactual or theory-based impact assessment. On the face of it, this could look surprising given the high expressed demand for complex methods (60% of surveyed civil servants argued that robust evidence (e.g. cost/benefit analysis, counterfactual analysis) is the

main criteria for judging the quality of recommendations. The apparent paradox can be explained by the types of questions addressed in evaluations. The most frequently used methods for data collection and analysis generally suffice for answering standard evaluative questions, which dominated the Lithuanian evaluations. The complex research designs and methods are necessary for answering causal questions that were not as widespread.

Type of evidence

Assessment of the types of evidence the evaluations generate – data, information or knowledge – is not straightforward, given the intangible nature of these types of evidence. To do so, we organised a focus group discussion with the members of the Evaluation Coordination Group. Overall, the discussants argued that evaluations deliver information, that is, sets of organised, transformed and analysed data that reflect the relevant dimensions of reality against evaluation criteria (e.g. relevance, effectiveness, efficiency). Participants also noted several evaluations that delivered knowledge. Such evaluations aimed to answer causal questions on the impact of interventions, relied on advanced methodological designs and aimed to generalise beyond the directly observed cases.

Use of the results of evaluations

We cannot directly observe whether and how the evaluations are used. Therefore, our analysis relies on two proxy measurements: (1) the types of recommendations produced by evaluations and (2) subjective perceptions of the surveyed civil servants. The meta-analysis included 675 recommendations from 57 evaluations published in 2007–2014. Each recommendation was classified by the type of impact its implementation would produce. Analysis included only those recommendations that were deemed as 'acceptable' by the relevant unit of public organisations, that is, the respective units expressed the intention to implement the recommendation. The meta-analysis revealed that most recommendations (42.5%) focused on improvement/fine-tuning of implementation of existing measures and instruments. Recommendations advocating change in intervention logic and policy priorities as well as improvements in monitoring and evaluation systems rank up next at 24 per cent and 22 per cent, respectively (see Figure 6).

The findings of the meta-analysis are supported by the results of a survey of civil servants, which included a question on the impacts of evaluations. The largest share of respondents argued that the evaluations helped to improve implementation as well as the monitoring and evaluation systems (see Figure 7). In addition, 14 per cent and 10 per cent of respondents argued that evaluations contributed to change in policy instruments and priorities, respectively. Finally, 17 per cent of respondents signalled symbolic use, by arguing that evaluations facilitated accountability.

Overall, both sources indicate that evaluations were largely geared towards improvements in implementation as well as monitoring evaluation systems. This suggests that evaluations are largely used instrumentally, although some evaluations are used conceptually (e.g. to review the logic of intervention) and/or symbolically (e.g. to report to the European Commission).

Conclusion and discussion

From the perspective of the models developed by Højlund (2014a) and Raimondo (2018), the Lithuanian evaluation system posed an interesting puzzle. It was set up due to pressure from the

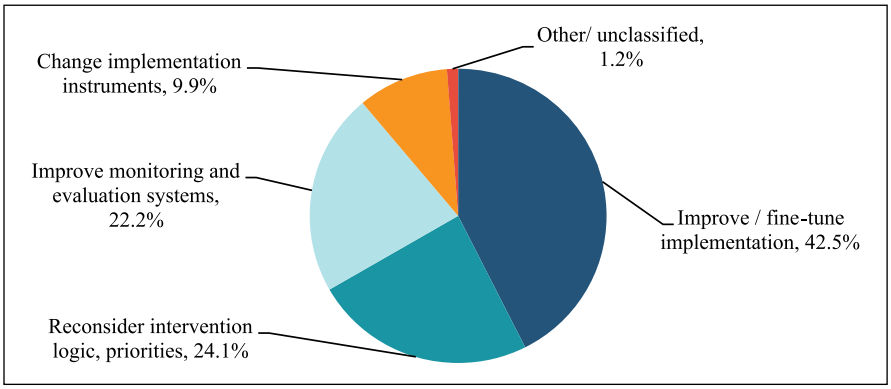


Figure 6. Recommendations by type.
N = 675 recommendations.

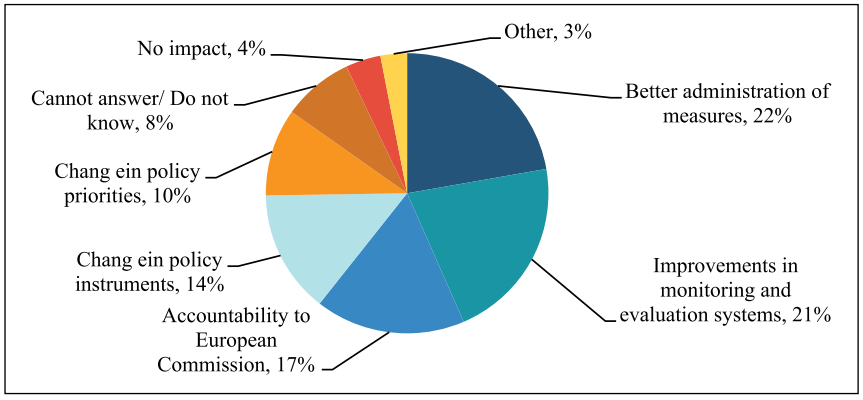


Figure 7. What was the impact of evaluations: the perceptions of the surveyed civil servants?
N = 60 civil servants.

EU to continuously evaluate the ESIF and it was embedded within post-communist legalistic administrative culture, and yet it produced evaluations that were used to improve the implementation and management of interventions. This article argues that the answer to the puzzle lies in the inner workings of evaluation systems. Evaluations are predominantly 'owned' by units of public organisations responsible for policy implementation and management of funds. Most evaluations in Lithuania deal with standard evaluative questions and provide information that can be instrumentally used to improve implementation, management and reporting systems – the primary interest of the dominant group of 'owners'. The characteristics of the five elements – (1) types of 'owners', (2) questions, (3) methodological designs, (4) type of evidence produced and (5) how it is used – are largely consistent and form a tightly knit evaluation system.

How useful are the 'ideal' types of evaluation systems as heuristic instrument? After all, the reality of evaluation systems is considerably more complex than the typology assumes. The evaluation steering groups are composed of diverse 'owners', the evaluation questions do not necessarily reflect their information needs and the evidence produced can (at least in principle) be used for a range of purposes. Hence, no real-world evaluation system will ever perfectly fit any of the 'ideal' types.

However, the purpose of heuristic instruments is to assist understanding of the characteristics of social phenomenon rather than describe it with large empirical accuracy. The proposed typology of evaluation systems has several benefits. First, it allows for a comparison of evaluation systems as well as tracking changes within an evaluation system over time. Anecdotal evidence suggests that during the early phases of development, the Lithuanian evaluation system produced largely data which was used symbolically. However, over time the system evolved and addressed initial limitations. Currently, it aims to address the challenge of a better integration of evaluative knowledge in policy design. Second, it allows moving beyond the dichotomy of good or dysfunctional evaluation systems. According to its own internal logic, each 'ideal' type is fit for a purpose, that is, promotes production and use of evidence in a way that satisfies the needs of its 'owners'. Third, the model outlines what outcomes the different elements of evaluation systems produce. This can assist the reforms aimed at strengthening the role of evaluation in the policy cycle.

The proposed approach to analysing evaluation systems also suffers from three important limitations that should be addressed by further research. First, it does not sufficiently account for the role of the administrative culture and normative frameworks that were highlighted by Raimondo (2018). Second, analysis of evaluation systems should be better embedded within the broader knowledge and performance management systems. This could provide a better understanding of the resources, motivations and value-dispositions of key actors. Finally, there is a need for better operationalisation of key elements as well as improved strategies for empirical measurement. The case study of the Lithuanian evaluation system mostly relied on proxies and the perception of key stakeholders – this provides 'soft' evidence at best.

Acknowledgements

We would like to thank Danutė Burakienė and Vilija Šemetienė from the Lithuanian Ministry of Finance for enlightening discussions, useful comments and facilitation of data collection process.

Funding

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

References

Ackoff R (1989) From data to wisdom. *Journal of Applied Systems Analysis* 16: 3–9.

Alkin M and King JA (2016) The historical development of evaluation use. *American Journal of Evaluation* 37: 568–79.

Barker K (2007) The UK Research Assessment Exercise: The evolution of a national research evaluation system. *Research Evaluation* 16(1): 3–12.

Boddington A (1993) Research evaluation systems: Sources of policy information and evaluation push. *Research Evaluation* 3(3): 197–203.

Caffrey L and Munro E (2017) A systems approach to policy evaluation. *Evaluation* 23(4): 463–78.

Choo CW (1998) *The Knowing Organization*. New York: Oxford University Press.

Christie CA (2007) Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation* 28(1): 8–25.

Cousins JB and Leithwood KA (1986) Current empirical research on evaluation utilization. *Review of Educational Research* 56: 331–64.

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

EVALSED (2013) *The Resource for the Evaluation of Socio-Economic Development*. Luxembourg: EUR-OP.

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

Grover V and Davenport TH (2001) General perspectives on knowledge management: Fostering a research agenda. *Journal of Management Information Systems* 18(1): 5–21.

Højlund S (2014a) Evaluation use in evaluation systems – The case of the European Commission. *Evaluation* 20(4): 428–46.

Højlund S (2014b) Evaluation use in the organizational context – Changing focus to improve theory. *Evaluation* 20(1): 26–43.

Højlund S (2015) Evaluation in the European Commission: For accountability or learning? *European Journal of Risk Regulation* 6(1): 35–46.

Knorr KD (1977) Policymakers' use of social science knowledge: Symbolic or instrumental? In: Weiss C (ed.) *Using Social Research in Public Policy Making*. Lexington, MA: DC Health, 165–82.

Leeuw FL and Furubo JE (2008) Evaluation systems: What are they and why study them? *Evaluation* 14(2): 157–69.

Leviton LC and Hughes EFX (1981) Research on the utilization of evaluations: A review and synthesis. *Evaluation Review* 5(4): 525–48.

Lillis DA (2000) Towards a new science evaluation system for New Zealand. *Research Evaluation* 9(2): 145–50.

Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.

Pelz DC (1978) Some expanded perspectives on use of social science in public policy. In: Yinger JM and Cutler S (eds) *Major Social Issues: A Multidisciplinary View*. New York: Free Press, 346–57.

Raimondo E (2018) The power and dysfunctions of evaluation systems in international organizations. *Evaluation* 24(1): 26–41.

Rich RF (1977) Measuring knowledge utilization: Processes and outcomes. *Knowledge and Policy* 10: 11–24.

Rip A and van der Meulen BJR (1995) The patchwork of the Dutch evaluation system. *Research Evaluation* 5(1): 45–53.

Rowley J (2006) The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33(2): 163–80.

Sullivan H and Stewart M (2006) Who owns the theory of change? *Evaluation* 12(2): 179–99.

The Ministry of Finance of Lithuania, VPVI and ESTEP (2013) *Vertinimo galimybių stiprinimas Lietuvoje: Patirtis ir pamokos* [Evaluation capacity building activities in Lithuania: Experience and lessons]. Vilnius: S Logistika.

Toulemonde J and Bjornkilde T (2003) Building evaluation capacity: Experience and lessons in member states and acceding countries. In: *Fifth European conference evaluation of the structural funds*, Budapest, 26–27 June.

Woolcock M (2013) Using case studies to explore the external validity of 'complex' development interventions. *Evaluation* 19(3): 229–48.

Žilvinas Martinaitis is an associate professor at Vilnius University and a research manager at Visionary Analytics. His academic interests cover learning organisations, knowledge management, and skills of workers.

Aleksandr Christenko is a researcher at Visionary Analytics. He is highly experienced in applying different statistical and data mining models for different tasks, including impact assessment, comparative analysis, and prediction.

Lina Kraučiūnienė worked at Visionary Analytics on evaluation and institutional capacity building projects. Her fields of interests cover coaching and empowerment of project teams to streamline processes and deliver high quality outputs.