CHAPTER 30

Policy Evaluation in Europe

Valérie Pattyn, Stijn van Voorst, Ellen Mastenbroek and Claire A. Dunlop

Abstract Policy evaluations are increasingly considered a taken-for-granted prerequisite for a well-performing public sector. In this chapter, we address the question whether this view reflects the actual situation concerning evaluation capacity and culture in Europe. First, we reflect on the history of policy evaluation in Europe, by distinguishing between two 'waves of evaluation': the countries in Northwestern Europe that have conducted evaluations since the 1960s and the countries in the rest of Europe for which evaluation is a more recent phenomenon. Next, to illustrate the two waves of evaluation, we zoom in on three political systems that represent the national, regional, and international level in Europe: the United Kingdom (UK), Flanders (Belgium), and the EU. For each system, we map evaluation culture and capacity by analyzing six indicators. The chapter concludes with a reflection on current trends in evaluation research and possibilities for future research.

V. Pattyn (✉)
Institute of Public Administration, Leiden University, Leiden, The Netherlands
e-mail: v.e.pattyn@fgga.leidenuniv.nl

S. van Voorst · E. Mastenbroek
Radboud University, Nijmegen, The Netherlands
e-mail: s.vanvoorst@uvt.nl

E. Mastenbroek
e-mail: e.mastenbroek@fm.ru.nl

C.A. Dunlop
University of Exeter, Exeter, UK
e-mail: c.a.dunlop@exeter.ac.uk

E. Ongaro and S. van Thiel (eds.), *The Palgrave Handbook of Public Administration and Management in Europe*,
https://doi.org/10.1057/978-1-137-55269-3_30

30.1 Introduction

While the idea of using evidence to inform policy is not particularly novel, attention for it in Europe has peaked in recent decades (Nutley 2003). Whereas evidence-based policy discourse initially dominated a limited number of policy areas, such as health, education, and social policy, it is now almost common sense in the entire public sector (Dahler-Larsen 2012).

The call to anchor policies in evidence can be considered the result of at least three interlocking tendencies. First, the increased interest in the economy, efficiency, and effectiveness of public policies. The New Public Management (NPM) paradigm, which has dominated the public sector reform agenda in many Western democracies from the late 1980s, is closely linked to this trend. To encourage effectiveness and efficiency, NPM has led to an increased delegation of policy implementation to autonomous agencies. Their establishment has posed important questions concerning the feedback lines of information, with governments requiring increasing amounts of evidence about how their policies function in practice.

A second factor is that policy-making has become increasingly complex. Policy issues are more and more intertwined, both horizontally between policy sectors and vertically between different government layers. For the European Union (EU) specifically, this is evident from the fact that many policies are nowadays initiated at the EU level, transposed into national legislation and implemented at the local or regional level. In such complex situations, the need for scientific policy analysis increases significantly.

A third factor is the desire for enhanced social responsiveness and societal support by involving citizens in the policy process. This idea has been nurtured by the governance debates since the 1990s and the notions of accountability, transparency, consultation, and participation associated with it. The provision of evidence on how public money is actually spent has become an important tool to reduce the distance between public administration and civil society (Brans and Vancoppenolle 2005).

Parallel to the diffusion of the evidence-based policy discourse, attention for policy evaluation as one particular type of evidence has risen dramatically. In this chapter, and inspired by Scriven (1991, 139), we conceive policy evaluation as 'the scientific analysis of a certain policy or part of a policy, aimed at determining the merit or worth of the policy on the basis of certain criteria' (Pattyn 2014a, 44). We cover all types of evaluations, ex ante, ex durante, and ex post. Policy evaluations are increasingly considered as a taken-for-granted prerequisite for a well-performing public sector (Dahler-Larsen 2012). In this chapter, we address the question whether this taken-for-granted aspect reflects the actual situation concerning evaluation capacity and culture in Europe.

It would exceed the scope of the chapter to investigate evaluation capacity and culture throughout Europe. The chapter takes an alternative approach. In Sect. 30.2, we reflect on the history of evaluation practice in Europe. We

distinguish between two 'waves of evaluation': the countries in Northwestern Europe that have conducted policy evaluations since the 1960s and the countries in the rest of Europe, for which evaluation is a more recent phenomenon. Next, to illustrate the two waves of evaluation, we zoom in on three particular cases: the United Kingdom (UK), Flanders (Belgium), and the EU (Sect. 30.3). Finally, Sect. 30.4 reflects on current trends in evaluation research and possibilities for future research.

30.2 Waves of Policy Evaluation in Europe

The first wave of evaluation in Europe has its origins in the United States (US), where policy evaluations were used to analyze how states implemented federal policies (Stame 2008, 119). Triggered by the Great Society (1960s–1970s), evaluation requirements were structurally incorporated in new social programs and policies. European countries that were part of this first wave were mostly located in Northwestern Europe, and included the UK, Germany, and the Scandinavian countries (Derlien and Rist 2002, 442; Bachtler and Wren 2006, 149). After the World War II, these countries developed mixed economies, in which government interventions were viewed as necessary instruments to correct the disadvantages of open markets. Accordingly, the 1960s and onwards saw an increasing demand for empirical and statistical information on the functioning of policies, in order to plan future interventions appropriately. This, in turn, triggered a need for policy evaluations (Vedung 2010, 265). The growth of government budgets and the willingness of (generally progressive) governments during this era also contributed to the growing numbers of policy evaluations in these countries (Derlien and Rist 2002, 442–443). During the first wave of evaluation, the randomized experiment was considered the gold standard in evaluation research, and evaluation researchers were expected to remain completely neutral (Vedung 2010, 266).

Importantly, not all countries in Northwestern Europe were affected equally by the developments described above. For example, the Netherlands and Switzerland remained relatively unaffected, possibly because their societies were divided between religious or linguistic groups, and therefore relied on their consensus culture as an alternative resource prioritization mechanism (Stern 2009, 80). Generally speaking, the countries that developed evaluation cultures during the first wave of evaluation have maintained their status as frontrunners in the field of evaluation until today (Jacob et al. 2015, 25).

The second wave, which began during the 1990s, originated in the US and the UK and has changed the nature of policy evaluation in Northwestern Europe (Stame 2008, 125). However, it also moved evaluation beyond this region to countries in Central, Southern, and Eastern Europe, which had been relatively slow to develop evaluation cultures (Bachtler and Wren 2006, 149). Generally speaking, two drivers pushed evaluation during the second wave.

The first driver was the NPM paradigm. Broadly speaking, NPM refers to the ideal of making governments more efficient and businesslike (Vedung 2010, 270). Whereas in the first wave, evaluation was primarily needed to improve policy implementation and learning ('what works'), in the second wave, 'accountability' became the leitmotif for evaluations. Attention to accounting for results is inherent to the NPM paradigm. By delegating authority to independent agencies, NPM weakened traditional command-and-control systems and created the need for alternative forms of information, such as evaluations (Lynn 2006, 143). Hence, methods typically associated with the second wave of evaluation are benchmarking and quantitative performance measurement (Vedung 2010, 272; Derlien and Rist 2002, 438–443; Stame 2008, 131). In the managerial tradition of NPM evaluations, attention is primarily biased toward output measures, at the expense of outcome and impact measures (Derlien and Rist 2002).

The second driver behind the second wave was pressure from the EU (Derlien and Rist 2002, 445–446). After Greece, Spain, and Portugal joined the EU during the 1980s, a large share of the EU's budget was redirected to these countries, in particular when it came to agriculture and regional development (Bachtler and Wren 2006, 149). The reform of the EU's regional policy involved the obligation for receiving countries to conduct evaluations to prove that funds were well spent (see Sect. 30.3 below).

During its more recent enlargement rounds in Eastern Europe, the EU required countries to build evaluation capacity before their accession. Therefore, Eastern European countries tend to have their evaluation activities concentrated in national ministries, while the countries in Southern and Northwestern Europe usually have more decentralized evaluation systems (Stern 2009, 80–81). Another consequence of the pressure from the EU is that the evaluation activities of the second-wave countries are often concentrated in the policy fields funded by the EU (e.g. regional development), while the evaluation systems of the first-wave countries are more diverse (Stern 2009, 81).

30.3 Present-Day Evaluation Culture and Capacity in Europe: Three Case Studies

Having sketched the drivers of the different waves of evaluation diffusion in Europe, one can ask whether and how these legacies have affected present-day evaluation practice. In this section, we address this question by mapping the evaluation culture and evaluation capacity of three political systems: the UK, Flanders, and the EU. These cases provide an equal representation of the national, the regional, and the international level in Europe. Furthermore, the UK and Flanders are suitable cases because they are typical of the first and second wave, respectively, and the EU is worth attention because it was a key driver of evaluation during the second wave.

30.3.1 Measuring the Evaluation Culture and Capacity of Countries

Defining and measuring evaluation culture and evaluation capacity is a challenging undertaking. The meaning of the concepts is ambiguous and often contested (Loud 2014, 58). A literature review by De Peuter and Pattyn (2009) of 16 evaluation capacity and culture-related sources revealed no less than 251 different indicators associated with the two concepts. The literature concerning evaluation capacity building is a rather practical field, rife with studies of success and failure in particular cases. Evaluation capacity and evaluation capacity building (ECB) is inherently contextual, which makes it difficult to identify common indicators that hold true across different settings.

Confusingly, the literature on evaluation capacity sometimes distinguishes between evaluation demand, or the value attached to evaluations in a given system (a definition which is very close to evaluation culture), and evaluation supply-, i.e. the resources and strategies essential for conducting evaluations (Nielsen et al. 2011, 327; Jacob et al. 2015, 27). In this contribution, evaluation culture refers to a shared understanding of the importance, functions, and roles of evaluation. Evaluation capacity, in turn, refers to the resources and strategies used to realize evaluations (Loud, 2014, 58). As such, we follow the predominant conceptualizations in the evaluation literature.

While much has been written on evaluation culture and capacity at the organizational level (e.g. Labin et al. 2012; Pattyn 2014a), there is relatively little academic literature on the operationalization of these concepts. An exception is the *International Atlas of Evaluation*, a widely referenced volume developed in 2002 and updated in 2015 (Jacob et al. 2015) to map the evaluation culture of approximately 20 political systems (Ibid., 7; Derlien and Rist 2002). The *International Atlas* systematically measures, scores, and subsequently compares countries' evaluation maturity by means of nine indicators (Jacob et al. 2015, 8). In this chapter, we roughly proceed from the same set of indicators to map the present-day state of evaluation culture and capacity in our three cases. Neither Belgium (including Flanders), nor the EU were part of the 2015 edition of the Atlas, although the UK was. Unlike the assessment conducted in the Atlas, we will not quantify the indicators on a single scale, but rather provide descriptive, qualitative information.

Within the scope of this chapter, we prefer to remove three of the less tangible (and hence, less measurable) indicators from the model used by the *Atlas*: the existence of a discourse in evaluation; pluralism in evaluations; and the involvement of different disciplines in evaluation. Table 30.1 provides a list of the six indicators that we will systematically describe. We also indicate whether each aspect is, in our view, primarily related to evaluation culture or to evaluation capacity.

30.3.2 Case Study: The United Kingdom

When evaluation began in the UK in the 1970s, it was driven by two forces. The first was the increasing professionalization of public management, which

Table 30.1 Six indicators for evaluation culture and capacity

Evaluation culture or capacity	*Indicator*
Evaluation capacity	An evaluation society exists
	There are institutional arrangements in the government (executive branch) for conducting policy evaluations and disseminating results
	There are institutional arrangements in parliament for conducting policy evaluations and disseminating results
	Policy evaluations occur within the supreme audit institution
Evaluation culture	Policy evaluation takes place in many policy domains
	Policy evaluations do not just focus on inputs/outputs, but also on outcomes

Source Adapted from Jacob et al. (2015), pp. 10–11

manifested itself in the creation of the Central Policy Review Staff (CPRS) in 1970—an early think tank which aimed to offer strategic inputs on policy priorities to the Cabinet. The second, more powerful force was the desire to control public spending in challenging fiscal environments which culminated in an International Monetary Fund (IMF) bailout in 1976. The Program Analysis and Review (PAR) was the process designed to ensure that resource consumption was tracked in governmental departments and results were reported to the cabinet.

In the 1980s, the CPRS gave way to the Efficiency Unit and Financial Management Initiative (FMI) which emphasized the use of business-inspired techniques to generate value for money in government. As NPM reforms took hold, the architecture of an 'evaluative state' had begun to take form (Neave 1988). While processes for cost evaluation—both ex ante and ex post—were well established by the 1980s (HM Treasury 1988), it was not until the second half of 1990s that procedures were formalized and routinized with the publication of the so-called Green Book (HM Treasury 1997, 2003, 2011) and formal adoption of impact assessment into government policy-making in 1998 (see Parker 2016 for an overview).

The focus on policy evaluation was systematized and broadened in 1997 by the first Labour administration since 1979. Its 1999 White Paper—*Modernizing Government* (Cabinet Office 1999)—heralded the expansion of evaluation techniques where uncovering what works constituted a new 'third way' of doing politics where ideology is replaced with pragmatic thinking (Giddens 1998). Guided by an ambitious conceptualization of evaluation, government focused not only on the disciplining power of evaluation for resource allocation, but on the learning capacity it generates with the aspiration of piloting large-scale social policies before full implementation (Cabinet Office 1999).

After New Labour and in the post-financial crisis years, evaluation has, in some respects, shrunk back to concentrate on delivering value—evaluation

should be 'comprehensive but proportionate' (HM Treasury 2011, 1). Resource management is now increasingly handled through a combination of key evaluation moments along with a more routine use of cost-benefit technologies as a part of impact assessments. The challenge for the government is to link the ex ante appraisals—where costs are estimated—with ex post evaluation—where they are realized (NAO 2013). The spirit of New Labour pilots lives on however in select policy areas with the 'What Works' research network that is responsible for £200 billion worth of spending (NAO 2013, 26) and in the rise of 'nudges' which use behavioral economics to inform policy design (see John 2014 for an overview).

Thinking about capacity and cultural metrics of evaluation, the UK evaluation society (UKES) established in 1994 provides a hub for policy evaluators in government, academia and business and the voluntary/third sector. This mix is reflected in its governance structure, with government researchers well represented on the UKES council. A strong presence at the annual conference and provision of training events attests to a well-rooted evaluation culture in the UK. Beyond this, there are also specialist bodies for particular analytical groups. So, for example, the UK's Social Research Association (SRA) provides a more tailored meeting place for Government Social Researchers (GSRs).

The UK's institutional arrangements for conducting and disseminating evaluations are strong. Evaluation tasks are decentralized to departments and, where appropriate, devolved administrations. Government Analytical Services commission evaluations conduct impact assessments and routinely evaluate expenditure. Each department team is led by a Chief Researcher who acts as a research champion (Dunlop 2010). The standards adhered to are enshrined in the aforementioned Green Book and so-called Magenta Book (HM Treasury 2011) which offers best practice examples and guides analysts on the suitability of techniques for different evaluative questions.

The importance of the UK's constitutional features for evaluation is crystallized when we consider the arrangements in the Westminster Parliament for conducting evaluations. The UK's executive and legislature are closely entwined—the former is composed of members of the latter. The aim is to promote stability and efficiency in the operation of government and policies and programmes are perceived to be the property of ministers and, by extension, parliament. This interconnectedness, and the primacy of the executive, leaves UK evaluation vulnerable to political rather than managerial agendas (Gray and Jenkins 2002, 130). Parliament is rightly seen as a passive evaluator. On all but financial matters, evaluation happens through distinct select committees of MPs who can investigate particular policy issues. While such evaluation is reactive—it is usually part of committees of enquiry triggered by crises or high-profile policy failures—the select committees are able to call experts witnesses from inside the civil service and outside the government to provide written and oral statements. While it does not fit our definition of evaluation as a robust and systematic form of policy review, these are

well-respected arenas and their access to expert elites is a significant power (Brazier and Fox 2011).

Turning to the presence of a supreme audit institution, we see a stronger role for the UK Parliament. A Parliamentary body, the National Audit Office (NAO) (and its devolved equivalents) provides select committees with reports on policy programmes. To give a sense of the scale of the NAO's work, in 2014–2015, it submitted 49 Value for Money reports and audited 442 accounts of 344 organizations covering over £1 trillion. Efficiency gains for that year were estimated at £1.15 billion (NAO 2015). Evaluation in the UK reaches across the majority of policy domains at a cost of £44 million in 2010/11 and employing around 100 members of staff. Yet, there is wide variation in how departments plan and apply evaluations (NAO 2013).

Finally, considering the objectives of evaluation, the aforementioned Green Book emphasizes the importance of understanding outcomes by capturing data on cost-effectiveness (HM Treasury 1997, 2011, Chap. 4). Yet, between 2006 and 2012, just over a fifth (70 of 305) of government evaluations included cost-effectiveness data with 4 of 15 chief departmental analysts classifying cost-effectiveness evidence as poor (NAO 2013).

30.3.3 Case Study: Flanders

Flanders is a typical case of a region belonging to the second wave of evaluation. Both NPM and the EU have played a major incentive for the introduction of policy evaluation, although, and as mentioned before, one should be careful about making unambiguous causal claims about these factors. The large-scale public sector reforms that the Flemish administration implemented in 2006, coined *Beter Bestuurlijk Beleid* (BBB), were clearly modeled along NPM blueprints. In this reform framework, the importance of evaluation was recognized and new instruments and tasks concerning evaluation were introduced. The framework charged departments with the responsibility for policy supporting tasks, whereas agencies were expected to generate the input by means of relevant policy and managerial information for policy evaluation. Although the structuring principles of the reform were not applied in all policy sectors, the introduction of the framework was a major trigger for policy evaluation, especially for departments that were not active in evaluation prior to BBB (Pattyn 2014b). In addition to NPM, the relevance of the EU as a driver for policy evaluation in the Flemish public sector is worth mentioning. Compulsory evaluation requirements associated with EU trainings and manuals spurred the building of evaluation capacity and culture. Yet, research (Pattyn 2015) does not confirm a spillover to policy domains not receiving EU subsidies.

The question is whether evaluation capacity and culture have reached full maturity in the meantime. In line with the federalist state nature, evaluation societies in Belgium are organized at the regional level. Evaluators active within the Flemish government and academics conceived the creation of the BBB setting as a window of opportunity to establish an evaluation society.

Consequently, the Flemish Evaluation Platform (*Vlaams Evaluatie platform* [VEP]) was launched in 2007. The large number of members of VEP (±1000) confirms the wide interest and need for sharing expertise about public policy evaluations.

With BBB, institutional arrangements have improved, as efforts were made to anchor policy evaluation in the policy cycle. A common trend for the different governmental levels in Belgium is to embed evaluation requirements in legislation. At the Flemish regional level, a number of ex ante 'tests' have been introduced, such as the child effect report, the poverty test or checks on the financial implications of new decrees for local governments. In 2005, and inspired by international and EU practice, the regulatory impact assessment procedure (RIA) was introduced, providing a framework for some of the tests mentioned 'A RIA is a structured analysis of the intended objectives and the expected positive and negative effects of the planned regulation in comparison with alternatives' (Kenniscel Wetsmatiging 2006). Yet, various evaluations of the RIAs revealed substantial shortcomings in the applications in the Flemish public sector (Bourgeois 2008). Especially with regard to the formulation of policy alternatives, criticism can be noticed. In addition, in many instances, the RIA is composed after the actual policy decision.

Whereas institutional measures are taken to anchor policy evaluation within the executive, the diagnosis is less positive for the parliament. Flanders follows the international trend in this regard (Jacob et al. 2015). This may be a surprise, given the strong accountability role that policy evaluations can potentially play. No special committee has been assigned the public policy evaluation role, as is the case in some other parliaments (Speer et al. 2015). The current president of the Flemish Parliament is nonetheless aware of the importance of policy evaluation. At his request, in the Autumn of 2015, the Socio-Economic Advisory Council developed 10 scenarios that would contribute to reinforcing (ex post) decree evaluation in the Flemish Parliament (SERV 2015). The influence of this report remains to be seen.

As a constituent body of Parliament, the Court of Audit of Belgium has seen an extension of its duty by the law of 1998 so that it became competent for performance audits. Most of these performance audits can be considered as ex post evaluations, but with the nuance that the assessment of effectiveness and efficiency does merely concern the investigation of boundary conditions and does not directly address the causal attribution question. In addition, in the performance audits, the official policy objectives are not put into question (Put 2005; SERV 2015).

Although evaluation in Flanders may be relatively limited still, activities are well spread across policy domains. Empirical evidence is found in varying sources, such as policy (planning) documents referring to existing or planned evaluation procedures and activities, announcements of public tenders and evaluation reports published on governmental websites. Of course, we can discern leaders and laggards. Policy sectors such as education, labor, and environment have a longer tradition and broader experience with evaluation than other sectors, as such reflecting wider international trends (Pattyn 2014a).

In terms of evaluation content, we can observe a pluralism of evaluation questions. Yet, goal attainment outweighs all other criteria in terms of attention. In an October 2015 survey launched by the Study Center of the Flemish Administration (Verlet et al. 2015) and sent to the top-level civil servants, 90% of the respondents indicated to (almost) always request an evaluation of goal attainment in evaluation questions. Sixty-two percent in contrast reported to (almost) always ask for effectiveness questions. Given the challenges related to measuring impact or effectiveness, this (self-reported) figure is still relatively high. Efficiency questions, in turn, are usually requested in half (50%) of the evaluations.

Considering all indicators, one can conclude that the Flemish public sector has made a major leap forward in terms of evaluation capacity and culture building in the recent decade. Differences between policy sectors remain, however, and often reflect international trends. Evaluations in Flanders are predominantly conducted for the sake of policy preparation and implementation, and the fulfillment of legal obligations, at the sacrifice of evaluations that are conducted for intrinsic policy learning purposes (Verlet et al. 2015). This observation confirms the legacy of the external drivers that characterized the second wave of evaluation.

30.3.4 Case Study: The EU

The EU's evaluation system has evolved over the past 30 years, which can be divided into four distinct periods (Højlund 2015). In the period 1980–1994, evaluation happened unsystematically in various DGs seeking to enhance policy effectiveness (Højlund 2015, 39). Typical forerunners were the spending DGs Development Aid, Science and Technology, and Regional Development (Summa and Toulemonde 2002, 409).

In the second stage (1995–1999), the increase of the structural funds and the advent of NPM resulted in a shift of the main evaluation motive from policy learning to financial accountability (Højlund 2015, 40). The number of evaluations of the EU Structural Funds Programs grew exponentially (Stern 2009, 79; Bachtler and Wren 2006, 146). Special guidelines and manuals (e.g. The Guide) were developed to assist member states in fulfilling their evaluation requirements. In addition, the Financial Regulation was amended to make evaluation of financial programs obligatory. Coordination of the evaluation system became the province of DG Budget.

The third period (2000–2006) started with extensive reforms after the resignation of the Santer Commission, aiming to enhance policy effectiveness and accountability. Ex post evaluation was broadened to include regulation and soft law. To this end, the DGs produced evaluation standards. Yet, despite the clear ex post focus of NPM, ex ante impact assessment became the key tool for enhancing effectiveness (Mastenbroek et al. 2015). The introduction of a highly advanced IA system diminished the incentive for carrying out ex post regulatory evaluation (Højlund 2015, 44).

The fourth period (2007-present) started with the introduction of the EU's Smart Regulation program, which put the notion of evidence-based policy-making center stage. To fulfill its renewed pledge to step up ex post regulatory evaluation, the Secretariat-General took over coordination, revising evaluation guidelines and introducing the Fitness Check instrument, designed to evaluate entire policy areas (Højlund 2015, 45).

Analyzing the EU's system, it should be noted that it lacks a clearly identifiable evaluation society or community. A network of Commission officials coordinating evaluation and impact assessment activities exists (Stern 2009, 71), but the external actors who conduct most evaluations are not involved in this network. There is also the European Evaluation Society with the aim of 'promoting the theory, practice and utilization of high quality evaluation in Europe and beyond,' but its scope does not correspond directly with membership of the EU.

The institutional arrangements for evaluation within the Commission are rather strong. The EU combines a decentralized system of DG units bearing evaluation responsibility central rules, guidelines, and standards on evaluation planning and content (EC 2015, 257; Stern 2009, 70–71). Important recent institutional additions are the explicit links to the budgeting system and the formal subjection of the evaluation units to internal audits (Højlund 2015, 36). A general intention to evaluate all EU activities is combined with frequent use of evaluation clauses—81% of all legislative evaluations conducted during 2000–2012 was based on an evaluation clause (Mastenbroek et al. 2015).

The institutional arrangements in the European Parliament (EP) are much weaker. Over time, the EP has enhanced its position, by conducting its own evaluations and increasingly calling for evaluation of regulatory evaluation. At the same time, actual interest in and use of evaluations by the EP has been meager, and driven primarily by agenda-setting motives (Zwaan et al. 2016). Poptcheva (2013) argued that the EP is generally distrustful of the quality of IAs. At the same time, the EP has sought to institutionalize evaluation capacity by establishing a Directorate for Impact Assessment and European Value Added to monitor Commission IA and evaluation activity (Poptcheva 2013, 5).

The European Court of Auditors (CoA) prioritizes auditing sensu *strictu*. Recently, however, it has embarked upon conducting performance audits, examining the economy, efficiency, and effectiveness of EU spending (Stephenson 2015). Effectively, this development has contributed to a blurring of the distinction between evaluation and audit (Smismans 2015). Also, the CoA seems to have assumed some system responsibility, given its research into the IA system, which also included some issues of ex post evaluation (European Court of Auditors 2010, 42).

Concerning coverage, the initial dominance by spending DGs has diminished, in response to the popularity of evaluation clauses and the explicit attention to legislative evaluations (Mastenbroek et al. 2015; Fitzpatrick 2012). These are distributed relatively evenly over the DGs: three out of four DGs with the largest share of major regulations and directives (DG Health

and Consumers, DG Environment, and DG Internal Market) are also in the group of four DGs that published the largest number of legislative evaluations (Mastenbroek et al. 2015).

Concerning evaluation objects, finally, most reports were found to simply describe implementation instead of analyzing outcomes or effects (Summa and Toulemonde 2002, 423). In a recent meta-study of 216 ex post regulatory evaluations conducted by the European Commission between 2000 and 2012, Mastenbroek et al (2015) established that 52% of evaluations contain at least some 'product evaluation' elements (goal achievement, effectiveness, efficiency, and/or side effects).

In sum, while the EU's current evaluation system has matured significantly, it can be further improved. It is characterized by two important tensions. First, the introduction of IA as the key mode for enhancing effectiveness has led to the construction of a poorly aligned policy 'silo,' next to the traditional ex-post evaluation system (Smismans 2015[1]). The two systems are surrounded by different evaluation capacities, the interdependencies of which could be better exploited (Smismans 2015). Moreover, despite its official rhetoric, the Commission has not yet lived up to its promise to provide systematic, high-quality regulatory evaluations. This has been linked to the lack of evaluability (Summa and Toulemonde 2002, 423) and the lack of suitable methods for outcome evaluation in a Union of 28 member states (Fitzpatrick 2012). Yet, this methodological challenge seems to be only part of the story, given the fact that the advent of evaluation in the US was actually linked to its federal character (Stame 2008, 120).

At the risk of oversimplification, in Table 30.2, we summarize the extent of evaluation culture and evaluation capacity in our three cases. In terms of capacity, evaluation is most vigorously rooted in the UK, which reflects its long tradition. Flanders and the EU have taken important steps forward, especially by reinforcing evaluation capacity within the executive. The picture for evaluation culture is more homogeneous. Evaluation is no longer the prerogative of a selection of policy domains, but is widely practiced across the entire public sector. Variation across policy domains still exists, at least in the UK and Flanders. In all three cases, effectiveness studies are conducted, but only cover a small share of evaluation activity. Further maturing in evaluation culture is hence still possible in all three cases.

30.4 Trends, Possibilities, and Challenges

30.4.1 Trends in Evaluation Research

To conclude, we reflect on some current trends in academic work about evaluation and its implications for future research. We do not focus on detailed methodological discussions found in evaluation journals, but refer to broader trends related to evaluation culture and capacity. In our view, three trends stand out when approaching the current evaluation literature through this lens.

Table 30.2 Present-day evaluation culture and capacity in the UK, Flanders, and EU

Evaluation culture or capacity	*Indicator*	*UK*	*Flanders*	*EU*
Evaluation capacity	An evaluation society exists	+	+	±
	There are institutional arrangements in the government (executive branch) for conducting policy evaluations and disseminating results	+	±	+
	There are institutional arrangements in parliament for conducting policy evaluations and disseminating results	±	–	±
	Policy evaluations occur within the supreme audit institution	+	±	±
Evaluation culture	Policy evaluation takes place in many policy domains	+	+	+
	Policy evaluations do not just focus on inputs/outputs, but also on outcomes	±	±	±

Key: + Yes; – No; ± Mixed
Source Adapted from Jacob et al. (2015), pp. 10–11

First, there is a trend towards research about evaluations of policies other than single measures, projects, and programs. The EU is a primary example of this: as the previous section indicated, it has gradually moved from evaluations of spending programs to a broader view on ex post evaluation, including evaluations of regulatory instruments. This development has also sparked a broad range of academic research (e.g. Bussman 2010; Mastenbroek et al. 2015; Smismans 2015).

Second, there is a trend towards more research about stakeholder involvement in evaluations. Although this issue already existed in the 1970s (Vedung 2010, 268), it has gained prominence in the literature during the last decade. This is particularly true for the literature on evaluation use, where stakeholder involvement is now commonly included as an independent variable and is generally shown to be a powerful explanation for use (Johnson et al. 2009, 379, 389). Other discussions in the literature include the potential bias in evaluation results caused by involving stakeholders (Cullen et al. 2011, 349) and the different ways in which stakeholders can be used (Ibid., 359). Stakeholder involvement has also been studied in specific European contexts, such as government organizations in Flanders (Pattyn and Brans 2014) and legislative evaluations in the EU (Mastenbroek et al. 2015).

Third, there is a trend towards studying alternative forms of evaluation use. For a long time, the literature on evaluation use was focused almost entirely on the use of evaluation reports for instrumental, conceptual, and strategic purposes (Henry and Mark 2003, 294; Johnson et al. 2009, 378). More recently, however, academics have started to explore other types of evaluation use, such as the positive effects of an evaluation process on the functioning of organizations (Ibid., 378; Shaw and Campbell 2013), the conceptual 'enlightenment' use of evaluations (Weiss 1977), and the long-term

impacts evaluations have on improving policies and organizations (Johnson et al. 2009, 378; Szanyi et al. 2013, 57). Related to this, instead of using the concept 'evaluation use,' evaluation jargon has shifted towards the term 'evaluation influence' (Henry and Mark 2003).

30.4.2 Evaluation Research: Possibilities and Challenges

Based on the case studies of evaluation systems and the trends presented above, evaluation research in Europe is facing at least three challenges for the years ahead. In the first place, there is a need for large-scale comparative research on the institutionalization of evaluation (Jacob et al. 2015, 7). While there is a vast body of literature on prescriptive theories and models for evaluation, most of the evidence on how evaluation works in practice is anecdotal, which makes it difficult to draw lessons across contexts. The evaluation literature would benefit from a more systematic assessment of the amount of means which various countries invest in evaluation, the extent to which their evaluations are centralized and explanations which account for these differences (Jacob et al. 2015, 28). Specifically, the influence of the EU on the uniformity or divergence of various evaluation systems in Europe would benefit from more empirical research (Bachtler and Wren 2006, 150–151). Conducting large-scale research requires a clear distinction between evaluation, monitoring, and performance management, to which this chapter has hopefully contributed.

Second, more systematic research is needed on stakeholder involvement in evaluation processes. A survey among 1683 evaluators in the US and Canada showed that stakeholder involvement is among the top issues about which evaluators would like to see future research (Szanyi et al. 2013, 56). In particular, they are interested in research about the effect of stakeholder involvement on evaluation impacts and about the perceptions of stakeholders of methodological approaches (Ibid., 57). While no similar survey has been conducted in Europe, these questions appear equally relevant for this continent.

Finally, we would call for more longitudinal and comparative research on the legacy of the two waves of evaluation. In particular, the relationship between NPM and evaluation requires attention. Scholars (Furubo and Sandahl 2002; Vedung 2010, 272) seem to agree on the influence of NPM on performance measurement and accountability-driven evaluations, but are less certain about the catalyzing role of NPM on evaluations focused on learning. The same applies to the role of the EU in fostering evaluation capacity in the member states. Much discussion still occurs with regard to the EU's impact on evaluation activity in the long run. While there is some consensus about the encouraging role of the EU in quantitative terms, more doubts exist about its influence on evaluation quality (Schwab 2009).

Note

1. Two additional systems, audit and enforcement, have their own, somewhat related logics (see Stephenson 2015).

References

Bachtler, J., & Wren, C. (2006). The evaluation of EU cohesion policy: Research questions and policy challenges. *Regional Studies, 40,* 143–153.

Bourgeois, G. (2008). *Mededeling aan de Vlaamse Regering van 8 September 2008 betreffende de evaluatie van de toepassing van de reguleringsimpactanalyse (RIA) en van de compensatieregel voor administratieve lasten—uitvoering van de regeringsbeslissing van 15 December 2006.* VR 2008 1909 MED.0421.

Brans, M., & Vancoppenolle, D. (2005). Policy-making reforms and civil service: An exploration of agendas and consequences. In M. Painter & J. Pierre (Eds.), *Challenges to state policy capacity: Global trends and comparative perspectives* (pp. 164–184). Basingstoke: Palgrave Macmillan.

Brazier, A., & Fox, R. (2011). Reviewing select committee tasks and modes of operation. *Parliamentary Affairs, 64,* 354–369.

Bussmann, W. (2010). Evaluation of legislation: Skating on thin ice. *Evaluation, 16,* 279–293.

Cullen, A. E., Coryn, C. L., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluation. *American Journal of Evaluation, 32,* 345–361.

Dahler-Larsen, P. (2012). *The evaluation society.* Stanford: University Press.

De Peuter, B., & Pattyn, V. (2009). Evaluation capacity: enabler or exponent of evaluation culture? In A. Fouquet & L. Méasson (Eds.), *L'évaluation des politiques publiques en Europe: Cultures et Futurs Policy and programme evaluation in Europe: Cultures and Prospects* (pp. 133–142). Paris: l'Harmattan.

Derlien, H., & Rist, R. C. (2002). Conclusion: Policy evaluation in international comparison. In J .-E. Furubo, C. Rayand, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 439–455). New Brunswick: Transaction.

Dunlop, C. A. (2010). The temporal dimension of knowledge and the limits of policy appraisal. *Policy Sciences, 43,* 343–363.

European Commission. (2015). *Better regulation toolbox* [complement to SWD(2015)111]. Brussels: European Commission.

European Court of Auditors. (2010). *Impact assessments in the EU institutions: Do they support decision-making? (Special report No. 3).* Luxembourg: European Court of Auditors.

Fitzpatrick, T. (2012). Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation, 18,* 477–499.

Furubo, J. -E., & Sandahl, R. (2002). Introduction: A diffusion perspective on global developments in evaluation. In J. -E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *The international atlas of evaluation* (pp. 1–23). New Brunswick: Transaction.

Giddens, A. (1998). *The third way.* Oxford: Polity Press.

Gray, A., & Jenkins, B. (2002). Policy and program evaluation in the United Kingdom: A Reflective state. In J. -E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International atlas of evaluation: Comparative policy evaluation*, 9. New Brunswick: Transaction. 129–153.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24,* 293–314.

HM Treasury. (1988). *Policy evaluation: A guide for managers.* London: HMSO.

HM Treasury. (1997/2003/2011). *Appraisal and evaluation in central government—The green book.* London: HM Treasury.

HM Treasury. (2011). *The magenta book—Guidance for evaluators.* London: HM Treasury.

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

Jacob, S., Speer, S., & Furubo, J.-E. (2015). The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later. *Evaluation, 21,* 6–31.

John, P. (2014). Policy entrepreneurship in UK central government: The behavioural insights team and the use of randomized controlled trials. *Public Policy and Administration, 29*, 257–267.

Johnson, K., Greenseid, L. O., Toal, S. O., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30,* 377–410.

Kenniscel Wetsmatiging. (2006). *Richtlijnen voor de opmaak van een Regulerings Impact Analyse*. Brussel: Vlaamse Overheid.

Labin, S. N., Duffy, J. L., Meyers, D. C., Wandersman, A., & Lesesne, C. A. (2012). A research synthesis of the evaluation capacity building literature. *American Journal of Evaluation, 33,* 307–338.

Loud, M. L. (2014). Institutionalization and evaluation culture—interplay between the one and the other. In M. L. Loud & J. Mayne (Eds.), *Enhancing evaluation use: Insights from internal evaluation units* (pp. 55–82). London: Sage.

Lynn, L. E. (2006). *Public management: Old and new*. New York: Routledge.

Mastenbroek, E., Van Voorst, S., & Meuwese, A. (2015). Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy*. Electronic publication ahead of print. http://www.tandfonline.com/eprint/kPHKRG4wwrMCsamnGbDk/full#.Vi9L-0n6rTcs, (doi:10.1080/13501763.2015.1076874).

National Audit Office. (2013). *Evaluation in government*. London: NAO.

National Audit Office. (2015). *Annual report and accounts 2014–2015*. London: NAO.

Neave, G. (1988). On the cultivation of quality, efficiency and enterprise: An overview of recent trends in higher education 1986–1988. *European Journal of Education, 23,* 7–23.

Nielsen, S. B., Lemire, S., & Skov, M. (2011). Measuring evaluation capacity—results and implications of a Danish study. *American Journal of Evaluation, 32,* 324–344.

Nutley, S. (2003). *Bridging the policy/research divide. Reflections and lessons from the UK. Keynote paper presented at the National Institute of Governance Conference*. Australia: Canberra.

Office, Cabinet. (1999). *Professional policy making for the twenty first century*. London: Cabinet Office.

Parker, D. (2016). Enterprise and competition. In C. A. Dunlop & C. M. Radaelli (Eds.), *Handbook of regulatory impact assessment*. Cheltenham: Edward Elgar.

Pattyn, V. (2014a). *Policy evaluation (in)activity unraveled. A configurational analysis of the incidence, number, locus and quality of policy evaluations in the Flemish public sector* [Ph.D. Dissertation]. Leuven: KU Leuven.

Pattyn, V. (2014b). Why organisations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation: The International Journal of Theory, Research and Practice, 20,* 348–367.

Pattyn, V. (2015). Explaining Variance in Policy evaluation regularity. The case of the Flemisch public sector. *Public Management Review, 17,* 1475–1495.

Pattyn, V., & Brans, M. (2014). Explaining organisational variety in evaluation quality assurance. Which conditions matter? *International Journal of Public Administration, 37,* 363–375.

Poptcheva, E. M. (2013). *Library briefing. Policy and legislative evaluation in the EU.* Brussels: European Parliament.

Put, V. (2005). *Normen in performance audits van Rekenkamers. Een casestudy bij de Algemene Rekenkamer en het National Audit Office* [Ph.D. Dissertation]. Leuven: KU Leuven.

Schwab, O. (2009). Europeanisation of German evaluation culture? On the effect of obligatory evaluation of European Union funds in Germany. In A. Fouquet & L. Méasson (Eds.), *L'évaluation des politiques publiques en Europe. Cultures et futures* (pp. 115–123). Paris: l′ Harmattan.

Scriven, M. 1991. *Evaluation Thesaurus* (4th ed.) Newbury Park, CA: Sage.

SERV (Sociaal-Economische Raad van Vlaanderen). (2015). *Tien denksporen voor ex post decreetsevaluatie in en door het Vlaams Parlement.* Brussel: SERV.

Shaw, J., & Campbell, R. (2013). The "Process" of process use: Methods for longitudinal assessment in a multisite evaluation. *American journal of evaluation, 35,* 250–260.

Smismans, S. (2015).Policy evaluation in the EU: The challenges of linking Ex Ante and Ex Post Appraisal. *European journal of risk regulation, 6*(1), 6–26.

Speer, S., Pattyn, V., & De Peuter, B. (2015). The growing role of evaluation in plarliaments: Holding governments accountable. *International Review of Administrative Sciences, 81,* 37–57.

Stame, N. (2008). The European project, federalism and evaluation. *Evaluation, 14,* 117–140.

Stephenson, P. (2015). Reconciling audit and evaluation? The shift to performance and effectiveness at the European Court of Auditors. *European Journal of Risk Regulation, 1,* 79–89.

Stern, E. (2009). Evaluation policy in the European Union and its institutions, in W.M.K. Trochim, M. M. Mark, & L. J. Cooksy (Eds.), *Evaluation policy and evaluation practice: New directions for evaluation* (pp. 67–85). San Fransisco, CA: Jossey-Bass.

Summa, H., & Toulemonde, J. (2002). Evaluation in the European Union: Addressing complexity and ambiguity. In J. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 407–424). New Brunswick: Transaction.

Szanyi, M., Azzam, T., & Galen, M. (2013). Research on evaluation: A needs assessment. *The Canadian Journal of Program Evaluation, 27,* 39–64.

Vedung, E. (2010). Four waves of evaluation diffusion. *Evaluation, 16,* 263–277.

Verlet, D., Lemaître, J., & Carton, A. (2015). Beleidsevaluatie binnen de Vlaamse overheid. Een Overzicht van de resultaten uit de bevraging van de leidinggevenden. Presentatie Studiedag Vlaams Evaluatieplatform. 17/12/2015.

Weiss, C. H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy Analysis*, *3*(4), 531–545.

Zwaan, P., Van Voorst, S., & Mastenbroek, E. (2016). Ex-post regulatory evaluation in the European Union: Questioning the use of evaluations as instruments for accountability. *International Review of Administrative Sciences.* doi:10.1177/0020852315598389.