



Why do Ministers Ask for Policy Evaluation Studies? The Case of the Flemish Government

Valérie Pattyn · Bart De Peuter · Marleen Brans

Published online: 4 December 2019
© The Author(s) 2019

Abstract Policy evaluations can be set up for multiple purposes including accountability, policy learning and policy planning. The question is, however, how these purposes square with politics itself. To date, there is little knowledge on how government ministers present the rationale of evaluations. This article is the first to provide a diachronic study of discourse about evaluation purposes and encompass a wide range of policy fields. We present an analysis of evaluation announcements in so-called ministerial policy notes issued between 1999 and 2019 by the Flemish government in Belgium. The research fine-tunes available evidence on catalysts for conducting evaluations. The Flemish public sector turns out to be a strong case where New Public Management brought policy evaluation onto the agenda, but this has not resulted in a prominent focus on accountability-oriented evaluations. We further show that policy fields display different evaluation cultures, albeit more in terms of the volume of evaluation demand than in terms of preferences for particular evaluation purposes.

Keywords Evidence-informed policy · Policy evaluation · Political discourse · Qualitative content analysis · Belgium

V. Pattyn (✉)
The Hague, The Netherlands
E-Mail: v.e.pattyn@fgga.leidenuniv.nl

B. De Peuter
Bruxelles, Belgium

M. Brans
Leuven, Belgium

Warum initiieren Minister Evaluationen? Eine Fallstudie der flämischen Regierung in Belgien

Zusammenfassung Politikbewertungen werden aus verschiedenen Gründen durchgeführt, u. a. um Rechenschaft abzulegen, um zu lernen und um öffentliche Politiken zu planen. Aber in welchem Verhältnis stehen solche Evaluationszwecke zu ihrer politischen Umgebung? Aktuell existiert nur wenig Wissen darüber, wie Regierungsministerinnen und -minister ihre Motivation für Politikbewertungen ausdrücken. Dieser Beitrag stellt daher eine neue diachronische Studie zu Diskursen über Evaluationszwecke in einer Reihe von unterschiedlichen Politikfeldern vor. Er enthält eine Analyse von Evaluationsankündigungen in sog. ministeriellen Politiknotizen der flämischen Regierung in Belgien zwischen 1999 und 2019. Die Forschungsarbeit verfeinert durch diese Analyse die verfügbare Evidenz zu evaluationsbegünstigenden Faktoren. Sie stellt fest, dass die Öffentliche Reformverwaltung die Politikevaluation im öffentlichen Sektor in Flandern etabliert hat, was allerdings nicht zu einem größeren Fokus auf rechenschaftsorientierte Evaluationen geführt hat. Unterschiedliche Evaluationskulturen in verschiedenen Politikbereichen beruhen größtenteils auf unterschiedlichen Evaluationsnachfragen und -volumina und weniger auf Präferenzen für einzelne Evaluationsmotivationen.

Schlüsselwörter Evidenzbasierte Politik · Politikevaluation · Politischer Diskurs · Qualitative Inhaltsanalyse · Belgien

1 Introduction

The attention to evidence-informed policy making has reached peak levels in recent decades (Davies et al. 2000; Head 2015; Pawson et al. 2011). This also pertains to policy evaluations as a particular type of evidence that is commonly associated with the evidence-informed policy movement. Policy evaluations hold the potential of ‘motherhood and apple pie’ (Tilley and Laycock 2000, p. 13), as they can bring about social betterment. From an instrumentally rational perspective, policy evaluations can be set up for multiple purposes, including accountability, policy learning and policy improvement, and policy planning (Schoenefeld and Jordan 2019; Vedung 1997). The question is, however, how these purposes square with politics itself. While there is some evidence on evaluation demand by parliamentarians (Speer et al. 2015; Bundi 2016), there is little knowledge on how government ministers conceive of the evaluation function and how they present the rationale of evaluations. Does this rationale differ across policy fields? And do we see any clear differences across government terms? In this article, we tackle this issue by unravelling the political attention to and discourse about different policy evaluation purposes across time. The article is the first to provide a systematic diachronic analysis of discourse about evaluation purposes and encompass a wide range of policy fields.

We present an analysis of discourse about evaluation studies that are announced in ministerial policy notes (*beleidsnota*'s) issued between 1999 and 2019 by the Flemish regional government in Belgium and, as such, spanning four government terms.

In such policy notes, ministers outline the policy priorities for the given government term in their particular policy field. In the (complex setting of the) Belgian federation, the Flemish government is in charge of a wide range of community matters such as education, culture, sports, youth and media; and regional competences on issues such as agriculture and fisheries, work and social economy, and mobility and public works. Belgium (and its regions), in an international comparative perspective, can be depicted as a case of the so-called second wave of Western countries/regions where policy evaluation only emerged on the government agenda in the late 1990s (Pattyn et al. 2018). The evaluation culture in Belgium has clearly matured in the last two decades, however, and especially in the Flemish public sector. Anno 2019, evaluation is relatively strongly institutionalised. We refer, for instance, to the establishment of a Flemish evaluation association in 2007, explicit debates on evaluation within and by parliament, a growing number of references to evaluation in policy documents and coalition agreements, the Court of Audit shifting part of its audits to evaluation of policy results and an extending supply of training in evaluation (Pattyn and Peuter 2020). Yet, as the international peloton also seems to keep the pace of maturing (Jacob et al. 2015), Belgium, including the Flemish public sector, probably has not compensated for its slow start.

To explain the agenda setting of evaluation in ‘second-wave countries’ (or regions) compared to early adopters such as the UK, Sweden or the Netherlands, scholars have resorted to the difference between internal pressures (early adopters) and external pressures (second wave countries). As relevant examples of external pressure for evaluation, the trends of New Public Management (NPM) and international cooperation (the European Union, in particular) are commonly stated (Furubo et al. 2002). How evaluation discourse is affected by these trends is, however, unclear. By focusing on the Flemish government, the current study provides a valuable complement to the many studies that focus on early-adopting countries, and can fine-tune available evidence on such catalyst factors for the agenda setting of evaluations.

Our analysis takes the perspective that evaluation is a rational tool par excellence for informing policy decisions, and functional for identifying the most effective and efficient means to reach societal goals. Evaluations can indeed also have a strategic or symbolic-tactical role; for instance, to hide shortcomings or failures (Vedung 1997; Widmer and Neuenschwander 2004). In fact, all evaluations are to some extent conducted for strategic-tactical reasons, with policy evaluation being political by nature (Bovens et al. 2006; Weiss 1993). Especially in evaluation discourse, ministers may be tactical in highlighting a particular evaluation purpose. While keeping the strategic potential of all evaluations in mind, in this article, we follow the mainstream taxonomy of evaluation purposes, which conceives evaluations as a mainly rational tool.

In ‘Theoretical framework and hypotheses’, we detail our typology of evaluation purposes for evaluating policies and draw up a set of hypotheses based on systemic context factors and existing literature. In ‘Methodology’, we explain the methodological pathway followed to conduct the analysis. The results are presented in ‘Findings’, and in the ‘Conclusion’, we reflect on lessons learned.

2 Theoretical framework and hypotheses

As Chelimsky and Shadish (1997, p. 18) have stated, the motive behind evaluation studies is of utmost importance, as ‘the purpose of an evaluation conditions the use that can be expected of it’. For policy makers, evaluations can serve multiple goals. In literature, various classification schemes can be discerned. From a rational perspective, three main purposes typically return (Schoenefeld and Jordan 2019; Vedung 1997).

First, evaluation can help to account for results vis-à-vis stakeholders. From a social mechanism perspective (Bovens 2010), the accountability approach frames in a principal-agent logic: public sector organizations are expected to provide feedback about their functioning and the results of their policies. Steering and accountability relationships can vary widely: between government and citizen, between donor and recipient, and between central and local governments. Evaluations that are set up for accountability reasons provide information to allow decisions on program continuation, expansion, reduction or termination (Bundi 2016, 2018). Evaluations, from this angle, can also serve an important outward-facing function. Via performance measurement tools and evaluations, politicians can signal their commitment to achieving certain goals, which can be useful to generate political trust. This can, in turn, help to mobilise political support and bolster the credibility of politicians (Boswell 2018).

Secondly, evaluations can be set up to support the decision-making process in the planning stage of the policy cycle. A wide range of evaluation questions can be tackled in this regard. Policy plans can be assessed according to their scope or urgency. Evaluations can assess and compare different policy alternatives and, as such, facilitate the decision-making process or the coherence and consistency of policy be checked in the planning process. Also, the policy relevance can be the object of evaluation prior to deciding on the implementation. It is relevant to mention that the European Commission considers policy relevance as essentially the most important evaluation criterion in its Better Regulation Agenda.

Thirdly, one can identify evaluations that serve policy learning. With policy learning, we refer to evaluations that contribute to the fundamental question ‘what works?’, or to related questions that apply a different approach to causality (e.g. ‘what works for whom in what circumstances and how’; Stern et al. 2012). Evaluations can be a useful tool for basic knowledge generation and help increase the general understanding of reality (Vedung 1997). Weiss speaks about the enlightenment function of evaluation (Weiss 1977). In the same vein, evaluations can be conducted for improving policy implementation by linking policy targets with the internal management of responsible organisations. Evaluation set up for this reason can deal with questions such as ‘Has the policy implementation been conducted in an efficient way?’ ‘Have stakeholders been sufficiently involved during the implementation?’ and ‘Are we on scheme to meet our policy objectives?’ A certain policy measure can be modified and improved, without changing the actual policy objectives. Argyris (1976) would label this type of learning as ‘single-loop learning’, whereas the revision of policy goals and the logic underpinning a certain policy initiative is rather to be conceived as ‘double-loop learning’.

Given the variety of motives as to why policy evaluations can be established, it is clear that policy evaluations can be of relevance in every stage of the policy cycle. Considering the boom of evaluation practices worldwide, some scholars have come to the conclusion that evaluation has acquired a ‘virtually sacred’ status (Dahler-Larsen 2012, p. 3). The question, however, is how such a statement should be empirically qualified. How do politicians view the evaluation function and can we observe certain trends in this regard? As mentioned above, apart from evidence on evaluation demand in the parliamentary arena (Speer et al. 2015; Bundi 2016), there is no such research focusing on government ministers. In our study, we particularly analyse the influence of macro- (New Public Management; EU dynamics) and meso-level variables (policy field) on evaluation demand.

A first trend that has been important in setting policy evaluation in motion in Belgium is NPM. Belgium is a relatively late modernizer, and NPM was only implemented on a large scale in the Flemish sector with the introduction of a public sector reform operation in 2006, coined Better Administrative Policy (*Beter Bestuurlijk Beleid*). The framework decree officially accompanying this reform was issued in 2003. Although public management reforms are qualified, contingent and variegated across countries (Pollitt and Bouckaert 2004), the Flemish reform operation complied exactly with key characteristics of the NPM blueprint (Brans et al. 2006). As outlined in the seminal article by Hood (1991), NPM can be described by seven doctrinal components:

1. hands-on professional management,
2. explicit standards and measures of performance,
3. greater emphasis on output controls rather than processes,
4. decentralisation of the administration,
5. more competition and contracting,
6. private sector styles of (personnel) management,
7. more parsimony in resource use.

The reform operation in the Flemish public sector was clearly modelled along these seven principles (Pelgrims 2008). The reform trajectory was initiated following calls from Parliament and from the administration which asked for a more comprehensive screening and modernization. The key objective of the reforms was to evolve to a public administration that is more transparent and more responsive to trends such as individualization, the growth of a network society, rapid evolutions in the field of information and communication technology, and globalization. With the reform, the Flemish government aspired to introduce a more client-oriented approach, an efficient and effective service delivery, and increased legitimacy by transparent structures. This implied an orientation toward results and decisiveness. The reform was built upon a vision (Stroobants and Victor 2000) promoting a cultural change based on the concepts of primacy of politics and political responsibility (De Caluwé and Van Dooren 2013). Organisation-wise, agentification became the leading principle. It was decided to restructure the Flemish administration around a number of policy fields that were built up of the same components: a core department for policy making with several internal and external agencies for the implementation of policies. Since the reforms, departments are also officially charged with the evaluation

of policy implementation and, more particularly, the effectiveness of the instruments used and the relations between output and outcome. During the implementation of the reforms, NPM's structural principles have been abandoned in some policy fields (see below), but the core traits are still visible. The question is, however, to what extent the introduction of the large-scale NPM-driven reforms in the Flemish public sector has impacted the political interest for evaluation, also in the longer run. In line with scholars as Furubo and Sandahl (2002), we posit that:

H1a NPM has acted as a lever for evaluation practice, which will be apparent by a continuous increase in political announcements of evaluations since the reforms.

When it comes to the framing of evaluations that the minister has in mind, it seems logical to expect that:

H1b NPM had a major influence on the purposes underpinning political announcements of forthcoming evaluations. Accountability-oriented evaluations can be assumed to have gained importance since the implementation of the reforms.

Alongside NPM, international cooperation has been considered as a major external push for evaluation (Furubo et al. 2002; Schwab 2009). The EU structural funds, in particular, are said to have played a key role in this regard. Linked to the granting of social funds for human resources and employment, territorial rebalancing, social cohesion and rural development, countries/regions had to prove these funds were well spent through monitoring and evaluation (Stame 2003). In this regard, the EU developed special guidelines and manuals. While evidence is not conclusive on the qualitative impact of the EU, its quantitative impact is uncontested (Schwab 2009), which we expect to retrieve in the study of Flemish ministerial policy notes.

H2 Intergovernmental policy dynamics in the EU have fostered ministerial demand for evaluation in a member state such as Belgium (i.e. Flemish public sector).

Besides these trends, which can be considered as systemic context factors, evaluation history is best read as a story of sectoral trajectories, with particular policy fields having integrated evaluation at a different pace. As highlighted by Meyer and Stockmann (2007) or Barbier (2012), evaluation practices are shaped by institutions that seldom operate across various policy fields. Internationally, a policy field such as education has a strong evaluation culture, for instance, with many methods of policy evaluation created and developed specifically in this sector (Crabbé and Leroy 2008). Although the Flemish public sector was relatively late in adopting evaluation practices, we assume that:

H3a The number of political announcements of evaluation will strongly differ across policy fields. In policy fields where there is a longstanding tradition in evaluation worldwide, we can presumably find more references to evaluation compared to evaluation fields without such a tradition.

Recent cross-country research by Jacob et al. (2015) on the institutionalization of evaluation supports this assumption. This study did not include empirical evidence for Belgium though. Studies on the demand of parliamentarians for evaluation (Bundi 2018; Speer et al. 2015) also pointed at important policy-field differences.

The question is whether such differences hold when focusing on ministerial interest for evaluations.

Elaborating on this, it can also be speculated that:

H3b In newly created policy fields (such as sustainability), ministers will announce a relatively low number of evaluations. And when announcing evaluations, these will presumably be more planning oriented.

Whereas policy evaluations do not necessarily require well-equipped monitoring instruments, previous research has revealed that public agencies usually invest first in the development of such monitoring tools, prior to proceeding to policy evaluation research (Pattyn 2014; Schoenefeld et al., this issue). Accordingly, we expect to find only a limited number of references in newer fields. In other words, when announcing evaluations, these will be mainly planning oriented, in view of the development of new policy measures.

3 Methodology

Our analysis focuses on evaluation discourse as it is used in ministerial policy documents in Flanders (Belgium). Evaluation discourse is commonly considered one of the key indicators to measure the maturity of an evaluation culture in a particular country or region, just as the extent to which policy evaluations are conducted in various policy fields (Furubo et al. 2002; Jacob et al. 2015). We have analysed four series of ministerial policy notes (*beleidsnota*'s) that altogether span a period of no less than 20 years of Flemish policy between 1999 and 2019. At the beginning of each five-year government term, a minister needs to submit such a policy note to parliament, in which he/she outlines his/her main intentions for his/her government portfolio per policy field. Importantly, while government ministers have the chance to put their 'fingerprints' on the documents, the policy notes reflect (coalition) government consensus. They are to be conceived as the further operationalization of the government agreement. All proposals mentioned in the policy notes are hence also backed up by the government, in principle.

For the 20-year period mentioned, we have examined policy notes relating to all policy fields for which the Flemish government is competent. In Flanders, policy notes constitute an important communication tool and have considerable weight: their implementation is intensively monitored by civil society organisations and parliament. Next to the government agreement, they are the most important reference tool for the government administration to prioritize tasks. The notes are a typical vehicle to announce evaluation intentions in light of future decisions about the introduction, modification or termination of policy measures. Although our study focuses on discourse about planned evaluations—and not on the evaluations that are actually implemented—we emphasize that references to evaluation are not non-committal, as these are actively monitored by parliament and societal stakeholders, especially in the neo-corporatist system of Belgium with its relatively institutionalized landscape of advisory councils. Of course, we realize that decisions about policy evaluations can also be communicated via other channels (such as evalu-

Table 1 Examples of types of evaluation purposes

Types of policy purposes	Example of corresponding citations (own translation)
Policy planning	“The studies for the construction of the open tide dock will be continued. We will conduct a societal cost–benefit analysis of the different planning options” (Policy note Mobility and Public Works 2009–2014)
Policy learning	“In the municipalities [...] where the influx of migrants from Central and Eastern European Countries and the concentration of Roma is relatively large, and cohesion is under pressure, we will deploy stewards. This initiative will run until 31 August 2016, after which an evaluation will be conducted, on which basis I can decide whether to continue the initiative or change it” (Policy note Home Affairs 2014–2019)
Accountability	“On condition of getting a positive evaluation, we will prolong the position of energy consultants” (Policy note Energy 2009–2014)

ation clauses in legislative decrees), or can be taken ad hoc, following a certain crisis, for instance. Notwithstanding these possibilities, the policy notes give us an important indication of the most important evaluations that are being conducted in a certain government term, and especially of the trends in the underlying purposes underpinning them. Although the author is, in principle, the minister as a political representative, administrations do contribute directly or indirectly to the content by providing context information, elaborated menus of policy choices, or advice on challenges and priorities. The same can be true for the agenda setting of evaluation exercises.

Each of the policy documents has been critically reviewed in search of citations referring to policy evaluation studies that ministers are planning on policy initiatives of the forthcoming term. A checklist of 20 key terms helped to identify the relevant citations: we systematically reviewed all citations mentioning (the Dutch equivalent) any of the following terms: ‘evaluation’, ‘planning’, ‘monitoring’, ‘pilot’, ‘benchmarking’, ‘experiment’, ‘comparison’, ‘efficiency’, ‘effectiveness’, ‘improvement’, ‘research’, ‘impact’, ‘audit’, ‘analysis’, ‘follow-up’, ‘try-out’, ‘verify’ and their respective conjugations. Given the inconsistent use of evaluation-related terms (De Peuter and Pattyn 2009), we did not restrict the analysis to citations that explicitly mentioned the term ‘evaluation’, but also screened for other terms that could refer to evaluations without using the term. Starting from the longlist of citations that included one or several of the key terms, we conducted a content analysis and only kept the citations that indeed referred to a concrete evaluation study and that clearly mentioned a reason why the evaluation would be carried out. As such, excerpts dealing merely with monitoring and not with evaluation were not considered.

Inspired by (Scriven 1991), we applied the following definition of a policy evaluation:

Policy evaluation is a scientific analysis of a certain policy (or part of a policy), aimed at determining the merit or worth of the evaluand on the basis of certain criteria (such as sustainability, efficiency, effectiveness, etc.).

By proceeding from a broad list of key terms, we could guarantee an encompassing approach to handle the large amount of data. In a subsequent step, we assigned

the relevant citations to a particular type of evaluation purpose. To ensure inter-coder reliability, the analysis was conducted by three researchers, who compared and cross-checked the identification and classification of citations.

Again, while recognizing the possible strategic and tactical use of evaluations, we restricted our analysis to the identification of three categories of purposes (Table 1).

4 Findings

For each of the government terms, Fig. 1 mentions the number of citations associated with each of the three evaluation purposes: policy planning (PP), accountability (AC) and policy learning (PL). As mentioned, we omitted the citations that did not refer to a concrete evaluation study and the few cases that did not allow for an unambiguous categorisation. From the data, the general observation is that both the total number of references to evaluation and the distribution between types of purpose remained relatively stable across time. As to the volume of references, the period 2009–2014 is a notable exception with an increased total. True, merely focusing on the absolute number of evaluation citations does not do justice to differences in potential size and budget across evaluations. However, within the scope of our analysis, it was not possible to include such indicators, also since evaluation budgets are difficult to retrieve in the case of in-house evaluation.

When considering the relative distribution of evaluation purposes, one out of three planned evaluations can be linked to policy planning. The latter refer to evaluations that scrutinize a certain policy measure, that compare policy alternatives or that consider the relevance of a particular policy's initiatives. About two thirds of the citations can be associated with policy learning. The share of quotes revealing an accountability-oriented motive is considerably lower, ranging from 3–9% only.

Importantly, the general picture of the relative distribution also holds true for the individual policy fields: the relative proportion of evaluation purposes does not vary strongly across fields. This is an interesting observation, especially since different policy fields often tend to favour varying evaluation approaches (Speer 2012). These 'sectoral evaluation styles' seem, generally speaking, not to be strongly reflected in

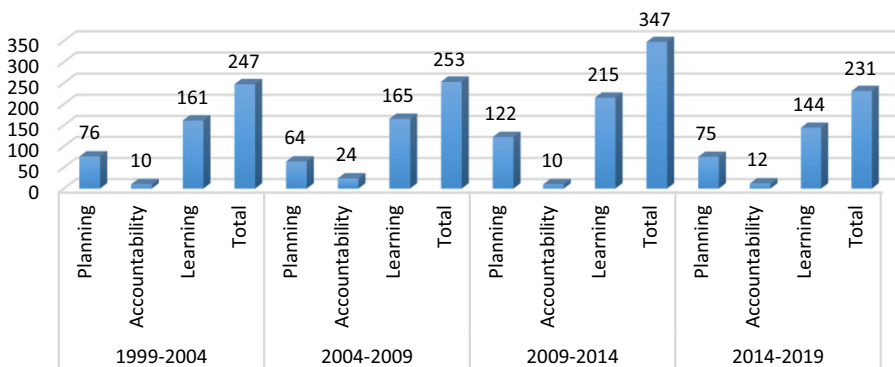


Fig. 1 Distribution of evaluation purposes across government terms

Table 2 Distribution of evaluation purposes across policy fields

Policy Field	PP	AC	PL	Total
Education	27	16	80	123
Environment (including Spatial Planning, as of 2014–2019)	39	4	75	118
(Social) Economy, Science, Technology and Innovation	31	7	60	98
Welfare, Public Health, Family, and Equal Opportunities ^b /Poverty ^c	32	3	58	93
Mobility and Public Works	50	5	33	88
Housing	13	2	38	53
Energy	18	1	29	48
Finance and Budget	16	1	30	47
Sports	8	3	24	35
Foreign Policy and International Development	12	1	19	32
Agriculture	10	0	21	31
Media	6	1	23	30
Culture	6	0	20	26
Planning and Statistics ^a , ^b /Horizontal Governmental Policy ^c , ^d	11	2	13	26
Home Affairs	4	0	16	20
Civil Service ^a /Administrative Affairs ^{b-d}	2	0	14	16
Tourism	4	0	12	16
Brussels Affairs	5	0	10	15
Youth	2	0	11	13
Total	338	56	685	1079

Policy fields that are in italics were later on subject of a separate policy note

PP Policy Planning, *AC* Accountability, *PL* Policy Learning

The superscripted letters refer to the policy notes' title for the first^a, second^b, third^c and fourth^d term of government

different preferences for specific purposes. Some fields nonetheless display a particular trend. We point, for instance, to the relatively high volume of planning-oriented evaluations in the field of mobility and public works. With a high burden on the budget, ex ante evaluations are fairly common practice in this field. One could also claim that the rather technical nature of the field lends itself relatively easily to ex ante studies.

This being said, the differences in the volume of references to evaluation between policy fields is more apparent. Table 2 lists those fields for which policy notes are available in all four government terms, to enable comparisons.

In the next section, we elaborate further on the hypotheses: we apply them to the case study of the Flemish public sector and verify to what extent they can be confirmed by the data. We also provide reflection on factors that can explain why some hypotheses appear more or less valid to this case. As stated earlier, we focus on the systemic trends referred to in literature as relevant triggers for evaluation culture and practice—NPM-inspired reforms and intergovernmental (EU) policy making—as well as on the comparison between policy fields.

4.1 NPM Public Sector Reform (Hypotheses 1a and 1b)

Has NPM acted as a lever for evaluation practice (Hypothesis 1a)? In our study, we conceive the establishment of the above-mentioned reform framework as the main manifestation of the Flemish government's adoption of NPM. As argued before, NPM is widely considered to have pressured laggards in formal evaluation cultures to adopt practices of formalisation and objectification, on which a policy-analytical culture could later build (Brans and Aubin 2017, p. 6). The strong increase in the number of citations in the period 2009–2014 seem to confirm the push that the NPM-inspired reforms gave to evaluation, yet with some delay. Although the reform framework was already implemented in the years before, research revealed (Pattyn 2014) that many government departments and agencies needed some time to prepare for their new tasks set by the reforms. This preparation process contributes to the understanding of why the number of citations in the two preceding legislatures—1999–2004 and 2004–2009—remained relatively stable. The decline in the number of citations for the most recent government term (2014–2019) corresponds with the loosening of the NPM reform principles, which may have contributed to the 'regression' of the volume of evaluation references. Therefore, we can conclude that hypothesis 1a can be confirmed by our data. As mentioned above, more than a decade after its introduction, it is now clear that the reform philosophy has not always been consistently implemented in practice (Fobé et al. 2017). In several policy fields, the evaluation capacity remained scattered between department and agencies. A few agencies are integrated in a department, and some policy fields have been merged.

Given the relatively late adoption of NPM in Belgium, the findings cannot be fully disconnected from the wider evidence-based movement. Belgian governments have embraced this global trend, both in discourse and practice: in addition to functions as forecasting and environmental analyses, evidence-based policy making also implies an investment in monitoring and evaluation (Fobé et al. 2017). While it can be argued that the general discourse on evidence-based policy gained momentum rather during the latest two of the four analysed legislatures, this is not reflected in our data. Importantly, there is no evidence that the EBP trend itself has slowed down in the Flemish public sector during the latest term of government. This suggests that NPM is probably more relevant to account for the decrease in evaluation references compared to the preceding period.

On the other hand, we see no confirmation in the data for hypothesis 1b—an expected shift towards accountability as a purpose for evaluation.

In fact, accountability continued to have a low share in the distribution of purposes. This brings us to the general conclusion that although the volume of evaluation references increased temporarily following the roll-out of the reform, there has not been a fundamental shift to a managerial logic of accounting for results, nor do we see a use of alternative frames since the implementation of NPM. In the entire period of investigation, a large majority of evaluation announcements are rather targeted at policy learning, with a share of 62–65%. These findings deviate from what has been found for the case of parliamentarians, where accountability turned out to be the major reason for demanding evaluations (Bundi 2016). The predominance of

the learning purpose may relate to a cumulative impact of two features: a ground layer of evidence-oriented attention in public debate and policy making which has steadily developed since the late 1990s on the one hand, and the fact that the learning purpose is logically connected to retrospective evaluation, which prevails.

Admittedly, our findings should also be conceived in light of the nature of our observation unit, i.e. policy notes as an important communication tool for a minister to announce policy plans for the coming term. Reference to policy evaluations in such a policy instrument can be considered as a minister's explicit intention to conduct an evidence-informed policy, to 'give account to' societal expectations in this regard. Although the accountability purpose can be expected to be implicitly present throughout the document, ministers can link other purposes (policy planning or policy learning) more directly to the policy decision-making process: decisions about the introduction of policy, the improvement of policy or policy termination.

4.2 Intergovernmental Relations Within the European Union (Hypothesis 2)

Alongside NPM, intergovernmental relations are regarded as important stimuli for launching evaluations, especially in these countries that were late adopters. Since the millennium change, a lot of policy fields have received direct questions or impulses to evaluate from the EU (Schwab 2009; Stame 2003; Speer 2012). Fields such as (social) economy (Stame 2003) and environment (Mickwitz 2013) are typical examples in this regard. In exchange for Structural Funds subsidies, the EU established a stringent system of evaluation requirements in these areas.

On the basis of our analysis, it is difficult to unambiguously link the trends in the number of citations to the impact of the EU across the different government terms in Flanders. Nonetheless, the relatively high number of evaluation announcements is no surprise. Ministers holding these policy fields in their portfolio already showed interest in evaluation prior to the NPM reforms and the evidence-based policy heydays in non-Anglo-Saxon countries, as the 1999–2014 data reveal. In the absence of another sound explanation, it seems safe to attribute this interest to a large extent to the EU evaluation requirements. We could not find robust indications of an increasing impact of EU cooperation on the evaluation volume across sectors. Thus, hypothesis 2—intergovernmental policy-making increases (demand for) evaluation—can only be partially confirmed. Importantly, the hypothesis holds true more for individual policy fields, rather than the total volume across policy fields.

4.3 Trajectories of Policy Fields (Hypotheses 3a and 3b)

A last dimension of analysis zooms in further on the comparison of policy fields. From the literature (Bundi 2018; Fobé et al. 2018; Barbier 2012; Speer 2012), we know that policy-field dynamics are relevant as they function as policy arenas on their own, characterized by specific sectoral identities and policy styles (Freeman 1985; Howlett 1991). This means that, all context factors being equal, policy-field dynamics also influence evaluation praxis. Also, our data point at an uneven distribution of references to evaluation across all policy fields. Some policy fields include dozens of references to citations that can be explicitly linked to an evaluation pur-

pose, while for others, we can only count a handful of references or no references at all. We draw particular attention to policy fields such as education; environment; (social) economy, science, technology and innovation; welfare, public health, family and equal opportunities; and mobility and public works, that excel in terms of a high number of citations. Altogether, almost half of the total number of citations (48%) can be assigned to these five policy fields. Two such fields, environment; and (social) economy, science, technology and innovation, have been named above as cases that receive substantial Structural Funds, which come with evaluation requirements. This does not apply to the other three fields ranking high: education; mobility and public works; welfare, public health, family and equal opportunities. In fact, the latter are also the fields that commonly top international evaluation maturity comparisons (see, e.g. Jacob et al. 2015). Flemish public administration ‘evaluation culture’ (Barbier 2012) thus seems to follow international trends. The findings also suggest that the minister holding a particular portfolio is of less importance than the policy fields he/she is heading (Schoenefeld et al., this issue), which resonates with previous research on parliamentary demand for evaluations in the case of Switzerland (Bundi 2018). We emphasize again, though, that the policy notes in Flanders reflect the consensus of the (coalition) government, making the analysis for individual characteristics of ministers, such as gender and political party, not very meaningful. Moreover, where differences exist between parties or gender, these can be attributed to the policy field a minister is heading. For instance, we indeed found a relatively larger volume of evaluation references for ministers of the Green Party compared to other parties. However, in the past 20 years, Green ministers were only in charge of three policy notes, of which two notes concern policy fields (environment; and welfare, public health, family and equal opportunities) which are among the fields generally displaying high evaluation demand, irrespective of the political party having the portfolio. Similar observations can be made for gender differences. Across the different government terms, female ministers tend to make more references to evaluation, generally speaking. Yet again, when taking their policy fields into account, differences between men and women are not outspoken. The small number of observations do not permit a more in-depth analysis, of evaluation purposes in particular, for these variables.¹

True, in some fields, the total number of citations fluctuates across different government terms. In (social) economy, science, technology and innovation, for instance, we find a much lower number for the 2014–2019 term compared to the previous government period, while for housing policy, this is the case for 2004–2009. Despite some irregular patterns, the overall distribution across sectors remains relatively stable across time. We can thus conclude that hypothesis 3a is confirmed by our data. Moreover, trends such as new public management reform or supranational cooperation do not seem to have a continuously growing effect on the volume of evaluation announcements in policy notes.

As to hypothesis 3b, we come to a mixed conclusion. Generally speaking, ‘newer’ policy fields (Table 3) only have a limited number of references to evaluation, which corroborates our expectations. The development of evaluation expertise is

¹ The data that support these findings are available from the corresponding author upon reasonable request.

Table 3 Distribution of evaluation purposes in newer policy fields

Policy field	1999–2004			2004–2009			2009–2014			2014–2019		
	PP	AC	PL	PP	AC	PL	PP	AC	PL	PP	AC	PL
Civic Inter- gration	–	–	–	0	0	2	0	0	4	0	0	0
Animal Af- fairs	–	–	–	–	–	–	–	–	–	0	0	1
Sustainable Development	–	–	–	0	0	0	–	–	–	–	–	–
Public–Private Cooperation	–	–	–	–	–	–	0	1	1	–	–	–

Cells containing a dash refer to government terms in which no policy notes were available in these fields
PP Policy Planning, *AC* Accountability, *PL* Policy Learning

a time-intensive undertaking, which might not be the priority of actors who are fully occupied with designing new policy measures. The analysis reveals that the Flemish government is focused on establishing monitoring equipment, rather than on fully fledged policy evaluation studies (Pattyn 2014). And when ministers do announce evaluations in such fields, we would logically expect that the purpose of policy planning is foregrounded. While the number of observations is too limited to draw strong conclusions, this expected bias is not shown in the data. The analysis is also constrained by the fact that a first self-standing policy note for a (new) field for a particular government term is not necessarily addressed separately in the next term. Even so, the observations in newer policy fields are symptomatic of the broader Flemish evaluation culture: also in policy fields with a strong evaluation maturity, ministers tend to be biased to ex post evaluations at the expense of ex ante evaluations (Fobé et al. 2017). Formal ex ante evaluations, which are directly linked to the planning purpose, are often restricted to the obligatory regulatory impact analyses. Further research will ideally verify whether this trend can be confirmed in other second-wave countries that were late in adopting evaluation practice.

5 Conclusion

Policy evaluation is an intrinsically political undertaking (Bovens et al. 2006; Weiss 1993). How ministers approach the evaluation function in public documents is largely a black box. In this article, we addressed this gap by analysing the volume and type of attention for different evaluation purposes in ministerial policy notes in which government priorities are outlined. With our focus on the Flemish government, the study provides an insightful longitudinal view of the agenda setting of policy evaluations in a region that has only recently adopted the evaluation practice, comparatively speaking. In evaluation literature, it is commonly assumed that several systemic drivers were influential in bringing policy evaluation onto the agenda. How such triggers have impacted the evaluation agenda and how they play out in the longer run is less known. The Flemish public sector turns out to be a strong case where NPM brought policy evaluation onto the agenda. The volume of

evaluation announcements largely follows the NPM dynamics in the public sector, generally speaking. When the attention to NPM wanes, so does ministerial interest in policy evaluation and vice versa. The findings do not point at an increasing impact of EU cooperation on the evaluation volume across sectors. As it has been shown in earlier studies, the history of policy evaluation has, to a large extent, developed along policy field lines (e.g. Barbier 2012). This sectoral pattern is clearly visible in the Flemish public sector, but mostly in terms of evaluation volume. Those policy fields excelling in evaluation maturity internationally speaking (and beyond the EU) are also the fields where we detect most evaluation announcements. As for explaining the type of evaluation purposes, our results are less conclusive. Contrary to our expectations, we could not retrieve a strong association of NPM with the announcement of accountability-oriented evaluations. Instead, we found a relatively strong dominance of announcements which are learning oriented. While one could argue that ministers will probably not be keen to initiate evaluations to be held accountable for the results of their policies, they neither seem to be extensively using evaluations for their outward-facing function (Boswell 2018), at least not in Flanders. Newer policy fields show no deviant pattern in this regard. Our expectation of finding more plans for evaluations focusing on policy planning in such fields cannot be confirmed.

For the administration in charge of the implementation of the evaluations announced in the policy notes and the evaluation community at large, our findings can be read as an incentive to engage in policy evaluations that are not primarily accountability focused, but that also enable policy learning. In fact, not all evaluation methods lend themselves to policy learning (Pattyn 2019). This is not to say, on the other hand, that parliamentarians cannot use learning-oriented evaluations to hold ministers accountable (Speer et al. 2015; Bundi 2016). Instead, they can verify to what extent actual evaluations are consistent with ministers' initial announcements.

The findings set the stage for a more extensive research agenda on this matter, which can address some of the limitations of the present study. Further research can allow fine-tuning of the interaction between the different factors and how they work in conjunction to set the evaluation agenda. Ideally, the quantitative approach to discourse analysis is complemented with a more qualitative outlook in which politicians are interviewed about their attitudes to the evaluation function. Such studies could provide more insights into the conditions under which ministers prioritize a certain evaluation purpose, and on the reasons why they strategically emphasize a certain purpose in policy discourse. Our research leaves the actual behavioural mechanisms unaddressed. In the same vein, research can engage with identifying the purposes of evaluations that are actually implemented by looking at the specific evaluation questions and the influence of evaluation findings on policy decisions. Is it indeed the case, that the evaluations that are learning oriented are eventually applied for these purposes? Finally, our findings apply to the Flemish case in particular. It would be interesting to compare our conclusions with a study of political attention to evaluation in other countries, to unravel divergence or convergence within and across different waves of evaluation practice.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution,

and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Argyris, Chris. 1976. Single-loop and double-loop models in research on decision making. *Administrative Science Quarterly* 21(3):363–375.
- Barbier, Jean-Claude (ed.). 2012. Introduction. In *Evaluation cultures. Sense-making in complex times*, 1–18. New Brunswick: Transaction Publishers.
- Boswell, Christina (ed.). 2018. Performance measurement and the production of trust. In *Manufacturing political trust: targets and performance indicators in public policy*, 5–29. Cambridge: University Press.
- Bovens, Mark. 2010. Two concepts of accountability: accountability as a virtue and as a mechanism. *West European Politics* 33(5):946–967.
- Bovens, Mark, Paul 'T Hart, and Sanneke Kuipers. 2006. The politics of policy evaluation. In *The Oxford handbook of public policy*, ed. Robert E. Goodin, Michael Moran, and Martin Rein, 319–335. Oxford: Oxford University Press.
- Brans, Marleen, and David Aubin. 2017. Introduction: policy analysis in Belgium—tradition, comparative features and trends. In *Policy analysis in Belgium*, ed. Marleen Brans, David Aubin, 1–10. Bristol: Policy Press.
- Brans, Marleen, Christian De Visscher, and Diederik Vancoppenolle. 2006. Administrative reform in Belgium: maintenance or modernisation? *West European Politics* 29(5):979–998.
- Bundi, Pirmin. 2016. What do we know about the demand for evaluation? Insights from the parliamentary arena. *American Journal of Evaluation* 37(4):522–541.
- Bundi, Pirmin. 2018. Varieties of accountability: how attributes of policy fields shape parliamentary oversight. *Governance* 31(1):163–183.
- Chelimsky, Eleanor, and William R. Shadish. 1997. *Evaluation for the 21st century: a handbook*. London: SAGE.
- Crabbé, Ann, and Pieter Leroy. 2008. *The handbook of environmental policy evaluation*. London: Routledge.
- Dahler-Larsen, Peter. 2012. *The evaluation society*. Stanford: Stanford University Press.
- Davies, Huw T. O., Sandra M. Nutley, and Peter C. Smith. 2000. Introducing evidence-based policy and practice in public services. In *What works? Evidence-based policy and practice in public services*, ed. Huw T. O. Davies, Sandra M. Nutley, and Peter C. Smith, 1–11. Bristol: The Policy Press.
- De Caluwé, Chiara, and Wouter Van Dooren. 2013. De Regionale Overheid: Vlaamse Overheid. In *Handboek bestuurskunde. Organisatie en werking van het openbaar bestuur*, ed. Annie Hondeghem, Wouter Van Dooren, Filip Derynck, Bram Verschuere, and Sophie Opdebeek, 161–190. Brugge: Vandenbroele.
- De Peuter, Bart, and Valerie Pattyn. 2009. Evaluation capacity: enabler or exponent of evaluation culture? In *L'évaluation des politiques publiques en Europe. Cultures et futures*, ed. Annie Fouquet, Ludovic Méasson, 133–144. Paris: l'Harmattan.
- Fobé, Ellen, Valérie Pattyn, Marleen Brans, and David Aubin. 2018. Policy analytical practice investigated: exploring sectoral patterns in use of policy analytical techniques. In *Policy capacity and governance. Assessing governmental competences and capabilities in theory and practice*, ed. Wu Xun, Michael Howlett, and M. Ramesh, 179–202. London: Palgrave Macmillan.
- Fobé, Ellen, Bart De Peuter, Maxime Jean Petit, and Valérie Pattyn. 2017. Analytical techniques in Belgian policy analysis. In *Policy analysis in Belgium*, ed. Marleen Brans, David Aubin, 35–56. Bristol: Policy Press.
- Freeman, Gary P. 1985. National styles and policy sectors: explaining structured variation. *Journal of Public Policy* 5(4):467–496.
- Furubo, Jan-Eric, and Sandahl, Rolf. 2002. Introduction. A diffusion perspective on global developments in evaluation. In *International Atlas of Evaluation*, ed. Jan-Eric Furubo, Ray C. Rist, and Rolf Sandahl, 1–23. New Jersey: Transaction publishers.
- Furubo, Jan-Eric, Ray Rist, and Rolf Sandahl. 2002. *International Atlas of Evaluation*. New Jersey: Transaction Publishers.
- Head, Brian. 2015. Toward more 'evidence-informed' policy making? *Public Administration Review* 76(3):472–484.

- Hood, Christopher. 1991. A public management for all seasons (in the UK)? *Public Administration* 69(1):3–19.
- Howlett, Michael. 1991. Policy instruments, policy styles, and policy implementation: national approaches to theories of instrument choice. *Policy Studies Journal* 19(2):1–21.
- Jacob, Steve, Sandra Speer, and Jan-Eric Furubo. 2015. The institutionalization of evaluation matters: updating the international atlas of evaluation 10 years later. *Evaluation* 21(1):6–31.
- Meyer, Wolfgang, and Reinhard Stockmann. 2007. Comment on the paper: an evaluation tree for europe. In *Conceptions of evaluation, rapport 08/2007*, ed. Gustav Jakob Perterson, Ove Karlsson Vestman, 139–148. Härnösand: NSHU.
- Mickwitz, Per. 2013. Policy evaluation. In *Environmental policy in the EU: actors, institutions and processes*, 3rd edn., ed. Andrew Jordan, Camilla Adelle, 267–286. London: Routledge.
- Pattyn, Valérie. 2014. Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3):348–367.
- Pattyn, Valérie. 2019. Towards appropriate impact evaluation methods. *The European Journal of Development Research* 31(2):174–179.
- Pattyn, Valérie, and Bart De Peuter. 2020. Belgium. In *The Institutionalisation of Evaluation in Europe*, ed. Reinhard Stockmann, Wolfgang Meyer, and Lena Taube. London: Palgrave Macmillan.
- Pattyn, Valérie, Stijn Van Voorst, Ellen Mastenbroek, and Claire A. Dunlop. 2018. Policy evaluation in Europe. In *The Palgrave handbook of public administration and management in Europe*, ed. Eduardo Ongaro, Sandra Van Thiel. London: Palgrave Macmillan.
- Pawson, Ray, Geoff Wong, and Lesley Owen. 2011. Known knowns, known unknowns, unknown unknowns: the predicament of evidence-based policy. *American Journal of Evaluation* 32(4):518–546.
- Pelgrims, Christophe. 2008. *Bestuurlijke hervormingen vanuit een politiek perspectief. Politieke actoren als stakeholders in Beter Bestuurlijk Beleid en de Copernicushervorming*. Brugge: Vandebroele.
- Pollitt, Christopher, and Geert Bouckaert. 2004. *Public management reform: a comparative analysis*. Oxford: Oxford University Press.
- Schoenefeld, Jonas J., and Andrew J. Jordan. 2019. Environmental policy evaluation in the EU: between learning, accountability and political opportunities? *Environmental Politics* 28(2):365–384.
- Schwab, Oliver. 2009. Europeanisation of German evaluation culture? On the effect of obligatory evaluation of European Union funds in Germany. In *L'évaluation des politiques publiques en Europe. Cultures et futures*, ed. Annie Foucquet, Ludovic Méasson, 115–123. Paris: l'Harmattan.
- Scriven, Michael. 1991. *Evaluation thesaurus*, 4th edn., Newbury Park: SAGE.
- Speer, Sandra. 2012. Sectoral evaluation cultures: a comparison of the education and labor market sectors in Germany. In *Evaluation cultures. Sense-making in complex times*, Vol. 1, 65–88. New Brunswick: Transaction Publishers.
- Speer, Sandra, Valérie Pattyn, and Bart De Peuter. 2015. The growing role of evaluation in parliaments: holding governments accountable? *International Review of Administrative Sciences* 81(1):37–57.
- Stame, Nicoletta. 2003. Evaluation and the policy context: the European experience. *Evaluation Journal of Australasia* 3(2):36–43.
- Stern, Elliot, Nicoletta Stame, John Mayne, Kim Forss, Rick Davies, and Barbara Befani. 2012. *Broadening the range of designs and methods for impact evaluations*. Department for International Development (DFID) Working Paper, Vol. 38. London: DFID.
- Stroobants, Eric, and Leo Victor. 2000. *Beter Bestuur. Een visie op een transparant organisatiemodel voor de Vlaamse administratie*. Brussel: Ministerie van de Vlaamse Gemeenschap.
- Tilley, Nick, and Gloria Laycock. 2000. Joining up research, policy and practice about crime. *Policy Studies* 21(3):213–227.
- Vedung, Evert. 1997. *Public policy and program evaluation*. London: Transaction Publishers.
- Weiss, Carol H. 1977. Research for policy's Sake: the enlightenment function of social research. *Policy Analysis* 3:531–545.
- Weiss, Carol H. 1993. Where politics and evaluation research meet. *American Journal of Evaluation* 14(1):93–106.
- Widmer, Thomas, and Peter Neuenschwander. 2004. Embedding evaluation in the Swiss federal administration: purpose, institutional design and utilization. *Evaluation* 10(4):388–409.