Polit Vierteljahresschr
https://doi.org/10.1007/s11615-019-00191-9

PVS Politische Vierteljahresschrift
German Political Science Quarterly

ABHANDLUNG

Following the Evidence Practice: An Analysis of Evaluation Studies on EU Railway Policy

Fabrizio De Francesco

Abstract Since the end of the 1990s, scholars have been paying particular attention to the link between evidence and policy because of the rise of evidence-based policy making and better regulation in the European Union political agenda. Documents such as evaluation studies are material traces of professional practice and the knowledge production process. Through the analytical perspective of evaluation practice, this contribution has two purposes. First, it differentiates three modes of evaluation theory and practice. Second, through a systematic content analysis of 52 evaluation studies of EU railway policy, it presents an overview of general patterns in the use of evaluative theories and practice. Besides contributing to the literature of evidence and policy practice, the article provides recommendations for EU evaluation and better regulation guidelines.

Keywords Content analysis of evaluation · Evaluation practice · European Commission · EU railway system · Knowledge production

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11615-019-00191-9) contains supplementary material, which is available to authorized users.

F. De Francesco (✉)
Glasgow, United Kingdom
E-Mail: fabrizio.de-francesco@strath.ac.uk

Published online: 22 November 2019

Evidenzpraxis in der Politik: Eine Analyse von Evaluationsstudien zur EU-Eisenbahnpolitik

Zusammenfassung Seit Ende der 1990er-Jahre beschäftigen sich WissenschaftlerInnen vermehrt mit dem Zusammenhang von Evidenz und Politik. Dieser Trend ist auf die wachsende Bedeutung von evidenzbasierter Policy-Gestaltung und auf die Agenda für Bessere Rechtsetzung (Better Regulation) der Europäischen Union zurückzuführen. Dokumente wie Evaluationsstudien sind materielle Produkte von Evaluationspraxis und des Wissensproduktionsprozesses. Ausgehend von einer analytischen Perspektive verfolgt dieser Artikel zwei Ziele: Zunächst unterscheidet er zwischen drei Arten von Evaluationstheorie und -praxis. Im zweiten Schritt werden allgemeine Muster der Verwendung von Evaluationstheorien und -praxis im Bereich der EU-Eisenbahnpolitik durch eine systematische Inhaltsanalyse von 53 themenrelevanten Evaluationsstudien dargestellt. Neben einem Beitrag zur Literatur der evidenzbasierten Politikgestaltung stellt dieser Artikel Empfehlungen für die Policy-Evaluation in der EU und für den Leitfaden für bessere Rechtsetzung zur Verfügung.

Schlüsselwörter Inhaltsanalyse von Evaluation · Evaluationspraxis · Europäische Kommission · EU-Eisenbahnsystem · Wissensproduktion

1 Introduction

Since the end of the 1990s, evidence-based policy making (EBPM) and better regulation (BR) have become the new mantras of policy making in Europe (Cairney 2016; Kirkpatrick and Parker 2007; Lee and Kirkpatrick 2006; Radaelli 2007, 2018). Each policy choice needs to be legitimated in terms of its economic rationality and scientific evidence (Liberatore and Funtowicz 2003). Therefore, within the European Union (EU) administrative process (Radaelli and Meuwese 2012), the Juncker Commission has strengthened the link between ex ante appraisal and ex post evaluation concerning both expenditure and regulatory policy (Smismans 2015a; 2015b). By closing the "evidence loop", the 2007 and 2015 EU BR reforms have enhanced the role of scientific evidence (Schrefler and Pelkmans 2014, p. 320) and risk management decisions in the formation of EU policies (Meads and Allio 2015). The renewed agenda also improved the independence of the regulatory oversight body, the Regulatory Scrutiny Board (RSB). With three external (to the European Commission) members, the RSB assesses the quality of impact assessments (IAs) as well as evaluations and "fitness checks" of existing legislation (Alemanno 2015, p. 351).

The resulting surge of administrative requirements for EBPM is characterised by the rise of evaluation as the practice of writing evidence. There is indeed a "metaphorical relationship between evidence and policy" that takes the form of evaluation practices (Freeman et al. 2011, p. 155) and is formally documented (Freeman and Maybin 2011). Evaluation is at the core of the EU governance (Smismans 2015b). Cohesion fund (Mendez and Bachtler 2011; Hoerner and Stephenson 2012), environmental (Bicchi 2011; Schoenefeld and Jordan 2017, 2019), and regulatory policy (Dunlop et al. 2012; Zwaan et al. 2016) are frequently evaluated by the European Commission.

The evolution of the evaluation system of the European Commission can be traced according to several functionalities (Højlund 2015) and waves of evaluation diffusion (Bachtler and Wren 2006; Johansson et al. 2015; Pattyn et al. 2018; Vedung 2010). The actual EBPM wave was preceded by the new public management (NPM) wave. Whereas financial control and accountability have been reinforced by the NPM movement, scientific methodology for obtaining generalisability of "what works" in a specific experience and context promoted the diffusion of EBPM. Each function and diffusion wave bring in distinctive evaluation prescriptions on the way to practice evaluation. Several evaluation approaches to EU policies have been proposed to categorise evaluation practices and methods (Pattyn et al. 2018; Stame 2008). By relying on the "evaluation theory tree" typology of methods, use, and valuing (Christie and Alkin 2008, 2013; Kallemeyn et al. 2015), this article aims to conceptually link the different functions of EU evaluation with actual practice as documented in evaluation studies conducted for the European Commission by external consultancies.

The starting analytical assumption is that evidence is produced in writing: "When evidence informs policy, the findings and conclusions of research and the problems and purposes of policy are distilled into documents" (Freeman and Maybin 2011, p. 155). Further, evaluation reports epitomise the theory and practice of evaluation. Therefore, evaluation studies can be categorised according to the different prescriptive assumptions and can be systematically analysed.

Within the EU governance, the policy-making process starts with evaluation studies contracted out by the European Commission to consultancy firms. No less than 80% of evaluations are conducted by external parties (Højlund 2014, 2015), generating a substantial part of the EU policy making. Although consultancies are a relevant actor in the EU evaluation process, external evaluation studies have rarely been analysed in a systematic way (but see Arnold et al. 2005; Cunningham 1997). To fill this gap, I maintain that shifts in the EU evaluation approaches should be reflected in external evaluation studies. The following question guides this paper: Is there a substantial variation across time of practice of evaluation as reflected in external evaluation studies conducted for the European Commission? To answer this question, this article focuses on the railway policy and systematically reviews 52 evaluation studies conducted for the Commission Directorate General for Mobility (DG MOVE) between 1999 and 2017. Although labelled as evaluation studies, these documents cover several aspects of railways such as market analysis, environmental and noise issues, and passenger safety. The wide scope of evaluation requires different sources of knowledge from professional policy evaluators, as well as from scientists, practitioners, and stakeholders.

The European Commission's strategy to create a single European railway market is a relevant case study for analysing knowledge production within the European Commission. Among utilities, the European railway system previously relied on country-specific rules and standards. Accordingly, the EU railway strategy has been characterised by a long and multistaged process of technical harmonisation and market liberalisation (Dyrhauge 2013; Finger and Messulam 2015; Nash 2008). The length of this process was also due to the fact that EU railway has been a highly political contested policy (Dyrhauge 2013). The European Commission summoned different types of external knowledge and technical expertise for taming the EU member states' resistance to liberalization. The empirical findings show that the evaluation theory tree provides a useful framework for differentiating the EU evaluation practice. Further, although the EBPM and NPM types of evaluation tend not to differ substantially across time, IAs rely mainly on EBPM, whereas NPM-oriented studies are mainly used in market as well as transposition and compliance analyses.

The remainder of this article is structured as follows. Section 2 provides an overview of the evolution of EU policy evaluation in the broader context of waves of diffusion practice. Section 3 establishes the link between evidence and evaluation practice. Section 4 presents the methodology and analytical framework that guide the systematic content analysis of external evaluation studies. Section 5 presents the empirical findings, and Sect. 6 concludes by providing policy recommendations for the EU evaluation system.

2 Unravelling the Evolution of Evaluation Practice in the EU

Evaluation has been the main accountability mechanism used by the European Commission to prove the transparency and efficiency of its policies (Stame 2008, p. 123) and, ultimately, achieve democratic accountability vis-à-vis other EU institutions (Højlund 2015). Throughout the growing process of Europeanisation, the European Commission's evaluation system has gone through several functional shifts (Højlund 2015). The current approach to evaluation is applied to all EU policies and bridges the conceptual and institutional divide between ex ante appraisal (mainly conducted on regulatory proposals through IA methodologies and cost–benefit analysis) and ex post evaluation for enhancing financial accountability and auditing (Smismans 2015b).

The emphasis on financial accountability was the remarkable element of the neoliberal wave and signified the consolidation of the NPM movement (Bachtler and Wren 2006; Højlund 2015; Pattyn et al. 2018). During this wave of confidence in customer orientation and market efficiency, evaluation materialised as an accountability mechanism for achieving economic efficiency and cost-effectiveness (Johansson et al. 2015; Vedung 2010). In the 1990s, value-for-money auditing was dominant, and evaluation has taken the form of performance measurement and consumer satisfaction through mechanisms of quality assurance and benchmarking (Bachtler and Wren 2006). This focus on financial accountability and the value-for-money approach has also been observed in the EU BR agenda (Radaelli 2007). In several EU member states, BR policy and (regulatory) IA were part of a broader political attempt to modernise public administration and enhance the competitiveness of a given country (Radaelli 2010). The political nature of IA was coherent with the NPM image of regulatory governance performance and accountability (Radaelli and De Francesco 2007; Radaelli 2010).

With the introduction of the 2007 EU Smart Regulation programme and the successive 2015 BR package, the evidence wave has reached the Commission evaluation system (Højlund 2015; Pattyn et al. 2018). The re-emergence of scientific methods was due to “a cruel disappointment” at the extent of effective reform accomplished through the NPM movement (Lapsley 2009). Further, the constellation of wicked problems such as climate and technological change forced governments and the EU institutions to go beyond internal management problems (Pollitt 2015; Lægreid and Verhoest 2019). Wicked issues can be solved by attempting to establish cause-and-effect relationships and observing what works in different contexts.

Although the most preferred evidence is obtained by proving the causal relationship between the policy treatments and intervention effects (mainly through randomised controlled trials and meta-analyses of scientific findings [Verdun 2010]), what constitutes evidence in evaluation—defined here as “a process of systematic inquiry to provide sound information about the characteristics, activities, or outcomes of a program or policy for a valued purpose” (King and Stevahn 2013, p. 13; see also Sanderson 2002)—is broader than EBPM concept and practice (Head 2008). For instance, traditional auditing tools and the evaluative opinion of frontline practitioners and users are also sources of evidence and common evaluation practices, especially within the NPM wave (Johansson et al. 2015). This distinction in types, origin, and perceived quality of evidence is also reflected in different evaluation practices and professional communities. Ultimately, the research puzzle is to understand to what extent these differences are present in the everyday evaluation practice and knowledge production of the European Commission. Before turning to the empirical part of this paper, the next section sets the argument that the production of evidence is a professional practice that can be categorised according to the evaluator’s epistemological position.

3 Evaluation as a Professional Practice

As Freeman, Griggs, and Boaz simply put it: “Evidence and policy are unthinkable without a concept of practice”. Practice refers to three connected elements: action, norms, and knowledge (Freeman et al. 2011). Practice denotes a process of action and has both material and social constituents. The first constituent refers to the intrinsic relation of entailment between actions and “artefacts”: Practice is developed through objects, tools, and instruments. And such artefacts embody practice (Freeman et al. 2011, p. 128). The second constituent is about the social dimension of practice: “Practices are very often carried out with others, and by reference to norms and standards that others, both participant and non-participant, will recognise” (Freeman et al. 2011, p. 128). This normative element resembles professional knowledge as “[p]ractices are competent performances” (Adler and Pouliot 2011, p. 4). A “practitioner” is a professional, an expert who competently performs and reiterates patterns of action that “are socially developed through learning and training” (Adler and Pouliot 2011, p. 5; citing Corradi et al. 2010). This reiteration of practice brings about routines and regularities of individual, professional, and organisational behaviour over time and space (Adler and Pouliot 2011, p. 6). Accordingly, professional knowledge and competence of evaluation, the third constituent, presuppose interaction and collaboration with other practitioners (Freeman et al. 2011, p. 131) and are structured in socially organised contexts (Dahler-Larsen 2011).

In evaluation, practice refers to "the everyday work of doing evaluation, such as dealing with stakeholders, developing an evaluation plan, collecting evidence, communicating findings, and so on" (Kallemeyn et al. 2015, p. 341–342). There are two approaches used by scholars to make sense of evaluation practice. The other strand considers evaluation practice as professional competence and knowledge emphasises how social, institutional, and global contexts "shape evaluators' understandings of their everyday work doing evaluation" (Kallemeyn et al. 2015, p. 342). The other strand considers evaluation practice as professional competence and knowledge stemming from either technical and instrumental rationality (and the application of scientific methods and procedures to knowledge production) or "practical" judgment and prior knowledge derived from previous experiences of evaluators, clients, and stakeholders (Schwandt 1997, 2005). All in all, evaluation practice is determined by evaluators' epistemological positions and is informed by theory (Shadish et al. 1991), as theorists are also "carriers of the practice" (Strang and Meyer 1993, p. 499).

The perspective of professional practices to produce evaluation can encompass evidence and the theoretical underpinnings. Evaluation practices can be distinguished according to taxonomies. Specifically, the "evaluation theory tree" emphasises three major areas of theorising related to evaluation practice: (i) methods of knowledge construction, (ii) valuing (descriptive or prescriptive approaches to addressing stakeholder values), and (iii) use of evaluation (Christie and Alkin 2013; Kallemeyn et al. 2015). The evaluation theory tree has been used for reflecting on the scholarly debate on evaluation practice (Christie and Alkin 2008, 2013) and to disentangle patterns of evaluation practice in Europe and the United States as presented in the scientific articles (Kallemeyn et al. 2015).

4 A Methodology for Disentangling Patterns of Evaluation Practice

4.1 The Unit of Analysis: Documents as Evaluation Practice

Documents enable communication across space and time. Therefore, "[g]overnment is unthinkable, impracticable, not feasible, without documents" (Freeman and Maybin 2011, p. 155). Everyday policy making is permeated by documents. Application forms, action plans, guidelines, targets, and performance review reports compose the daily activities of policy makers and civil servants, and there is no policy without documents depicting, justifying, and legitimating the decision-making process (Freeman 2009; McGrath 2016). This is particularly true for evaluation. As objects of decision making, the creation of knowledge and evidence are made of documents, and documents facilitate the circulation of knowledge and evidence. Summarised in documents such as evaluation studies, empirical methods allow knowledge to "be categorized, quantified, tested, externally validated, and relied upon to provide answers" (Eversole 2012, p. 34). All in all, as policy documents, evaluation studies are not only material traces of actions, but they embody the knowledge production process. As a practised thing, evaluation studies are a medium to observe evaluators' actions and reasoning.

The methodological argument here is that similarly to policy actors, a decision-making process can be traced through the practice of recreating and reproducing documents as artefacts for summarising evidence and knowledge. The emphasis is on the instrumentality of documents (Prior 2008) in constituting and maintaining a specific discourse of evaluation practice revolving around the EBPM or NPM diffusion wave. This argument is based on the fact that documents are vital in creating and nurturing epistemic communities as well as communities of (evaluation) practice (Bicchi 2011).

4.2 Evaluation Theory Tree as an Analytical Framework

The analysis of documents as an evaluation practice relies on the categorisation of the evaluation theory tree. Evaluation theories are here appreciated for their prescriptive dimension and links to practice: "[T]hey offer a set of rules, prescriptions and prohibitions that specify what a good or effective evaluation study is and how an evaluation study should be conducted. None of the evaluation approaches is predictive or offers an empirical theory" (Christie and Alkin 2008, p. 132).

Alkin and Christie (2013, p. 12) depict the evaluation tree as a trunk with three main branches. The trunk is built on accountability and systematic social inquiry: "While accountability provides the rationale, it is primarily from social inquiry that evaluation models have been derived" (Alkin and Christie 2013, p. 12). The first column of Table 1 shows that the first branch of methods is the continuation of the social inquiry trunk. Because this method branch is about how to obtain generalisability or knowledge construction (Alkin and Christie 2013, p. 12), it is associated with EBPM and the evidence wave of evaluation. Accordingly, the evaluation practice is based on a careful discussion about methodology and the accuracy of data. Randomised controlled trials, experimental or quasi-experimental research design, and meta-analysis are preferred to other types of evidence and evaluation practice. The methods branch revolves around quantitative analysis in order to establish cause-and-effect relationships. Overall, scientific knowledge and scientific methodology are central to this branch. This is reflected in the characteristics of the documents. Evaluation studies of the methods branch contain formulas and statistical models. The predominance of science is also epitomised in citations of scientific articles and other evaluation studies: "Especially in science, the ability of a document to convince its readers crucially depends on the way in which that document 'enrolls' other documents as 'allies'" (Simons 2016, p. 53). Therefore, citations are instrumental for pursuit of knowledge authority. Because of the extensive reliance on scientific methodology, this type of analysis can be conducted for both ex ante appraisal and ex post evaluation.

The second branch, valuing, is associated with the theoretical and ideational trunk that considers evaluation about making judgments on the merit of public policy and programmes. The professional role of the evaluators is central in this evaluation branch as they are under demand to "place value on their findings and, in some cases, determine which outcomes to examine" (Christie and Alkin 2013, p. 32).

Table 1 Evaluation theory branches and evaluation features

Features	Evaluation branches		
	Methods	Valuing	Use
Social context and wave of diffusion	Evidence-based policy making	New public management and neoliberal wave	Dialogue-oriented wave (pluralistic model of evaluation)
Scope	Best scientific evidence; focus on understanding cause–effect of interventions	Making judgments; focus on accountability	Continuous organisational learning process; focus on dialogue and participation of stakeholders
Main goal	Scientist methodology	Cost-effectiveness	Stakeholders' view
Stage of evaluation	Ex ante and ex post	Ex post	Ex ante
Practice	Experiments and quasi-experiments; systematic literature review and meta-analysis	Indicators, performance measures, ranking and benchmarking; comparison and assessment of options	Surveys and process-oriented analysis; participation of stakeholders in evaluation; critical discussion of objectives; design of information system and process
Research method	Quantitative models for cause-and-effect relationships	Indicators and quantitative measures for ranking and assessing performances and policy options	Qualitative case studies and critical analysis
Document elements	Citations of scientific literature and previous evaluation studies. Formulas and statistical results	Comparative graphs and statistics	Graphs representing decision-making process

The value branch lies largely on the theoretical foundation of accountability. In order to ensure efficient and appropriate use of resources, public interest is the main evaluation criteria of this branch (Ryan 2004). By relying on their expertise and knowledge, evaluators need to compare competitive policy alternatives in order to make value judgments. These types of studies rely also on stakeholders as primary sources of data or judgment perspectives for evaluation. Indicators for assessing policy options are central in these types of studies. Within the NPM wave, indicators would score policies and programmes for their value-for-money, cost-effectiveness, and customer orientation. An NPM-style evaluation study revolves around the results of public intervention (through performance-based management, management by objectives), rather than discovery of the cause-and-effect relationship. It is characterised by an emphasis on ex post evaluation through performance measurement, consumer satisfaction appraisal and quality assurance, and benchmarking (Vedung 2010, pp. 271–273). Evaluation studies will contain comparative graphs and tables summarising the variation in performances.

The third branch, use, is composed of practices developed in order to make sure that decision makers and stakeholders actually utilise evaluation. The orientation of this prescriptive theory is towards the effect of evaluation on decision making

(Christie and Alkin 2013, p. 13). The prescriptions here regard the way evaluation will be used. The use branch focuses on stakeholders who will use the evaluation information (Christie and Alkin 2013, p. 13). Prescriptions of this theoretical branch are also about "how to find out and represent various stakeholder groups" (Ryan 2004, p. 444). In this dialogue-oriented way of conducting evaluation, typical of pluralistic models of evaluation, no interaction with practitioners and stakeholders would lead to policy failure. Use-oriented evaluation studies "are designed specifically to assist key program stakeholders in program decision making" (Christie and Alkin 2013, p. 44). This type of study is dedicated to decision-making processes that are often represented in graphs. Use-oriented evaluation studies also revolve around stakeholders' professional judgments. Stakeholders are constantly involved in the evaluation process. Surveys of stakeholders' views on policy and programmes is a common practice of the use evaluation branch. Use-oriented evaluation studies are practice-based case studies providing a qualitative discussion on programme objectives and the extent of achievement of such objectives. This theoretical approach is suitable for ex ante appraisals. This is because there is a political discussion on policy objectives, and evaluation should design a system for a continual information stream to decision makers in order to ensure that programmes and policies continually improve (Christie and Alkin 2013, p. 44).

4.3 Systematic Document Analysis as a Research Method

The relationship between knowledge expressed in documents and policy making is not novel in the literature of evaluation and public policy. Positivist scholars tend to focus on the substantive content of a document. For instance, the frequency of scientific citations in the United States government's regulatory impact analyses provides a proxy of the quality of EBPM (Desmarais and Hird 2014). Social constructivists are also interested in evaluation documents, but they focus on the language of the documents as text (Freeman and Maybin 2011, p. 157), as in the case of the European Commission's IAs that have been analysed for their narrative (Radaelli et al. 2013).

This article relies on a substantive analysis of evaluation studies' content and structure (Bowen 2009; Prior 2008). Following Prior (2008), this review focuses on both the different practices and the function of evaluation studies that characterise the different branches of evaluation theory and practice. Specifically, the content of evaluation studies is analysed for associating a single evaluation study with a specific evaluation branch. Graphs, research design, methods, models, and other evaluation practice features (Table 1) guided the categorisation into one of the three distinctive professional practices of evaluation.

Content and thematic analysis of evaluation reports is not novel in the assessment of the EU evaluation. Højlund (2014) included evaluation reports and IAs in his analysis of use of evaluation within the LIFE programme. Huitema et al. (2011) mapped the climate policy evaluation practice in Europe. In a similar vein, Jordan and Schoenefeld (2019) proposed a conceptual map for detecting complementarities and potential tensions between drivers of environmental policy evaluation. Torriti (2010) analysed an IA for assessing the quality of economic analysis concerning EU energy liberalisation. The methodological contribution here is to analyse the content and theme of evaluation studies as "the ways in which different identities, beliefs and values come to play a role in explanations of particular ways of doing things" (Freeman et al. 2011, p. 129).

4.4 Sample of Evaluation Studies

The documents included in this analysis are evaluation studies commissioned by DG Move (previously established within the broader DG Energy and Transport) to external consultancies between 1999 and 2017. The table presented in the online appendix summarises all consultancy reports downloaded from a DG Move web page dedicated to railway studies.[1] It presents the date, title, name of consultancy, and evaluation practice utilised in the studies. Turning to policy sectors, these evaluation studies are mainly on the technical harmonisation of the European rail system, the environmental impact of railways, market analysis, and transposition of EU legislation by EU member states. Finally, DG Move commissioned several studies for supporting the IAs of regulatory proposals and railway packages. Further, the table associates each study with an evaluation theory branch.

5 Empirical Findings of the Systematic Content Analysis

Although the first directive for liberalising the European railway market dates back to 1992, it was only at the beginning of the 2000s that the EU adopted a step-by-step strategy for opening and increasing the competitiveness of the railway market (Di Pietrantonio and Pelkmans 2004). Four reform packages were necessary for creating a single European railway area based on three principles of economic governance: (i) financial separation between rail infrastructure managers and providers of rail service (achieved through the 1991 directive and second package); (ii) interoperability and technical harmonisation of national rail systems (achieved through the second and third packages); and (iii) transparency of the licensing process through the establishment of national regulatory agencies and the European Railway Agency that is now acting as a centralised one-stop shop for licences and safety certification of rail operators (achieved through the second and fourth packages).

Fig. 1 shows that the larger yearly productions of consultancy reports coincided with the adoption of the railway reform packages in 2001, 2004, 2007, and 2016. In particular, the research activities leading to the 2004 reform were particularly relevant: Between 2002 and 2004, as well as between 2005 and 2007, the European Commission commissioned 14 and 12 evaluation studies, respectively.

The table in the online appendix lists the leading external consultancies that have been commissioned by the European Commission to conduct evaluation studies. Although the vast majority of consultancies produced only one evaluation study, there are two consultancies that have been commissioned more frequently: Steer Davies Gleave drafted 10 studies as the leading consultant, three of them being IAs,

[1] https://ec.europa.eu/transport/modes/rail/studies/rail_en, accessed on 22 Aug 2014 and 27 Dec 2017.

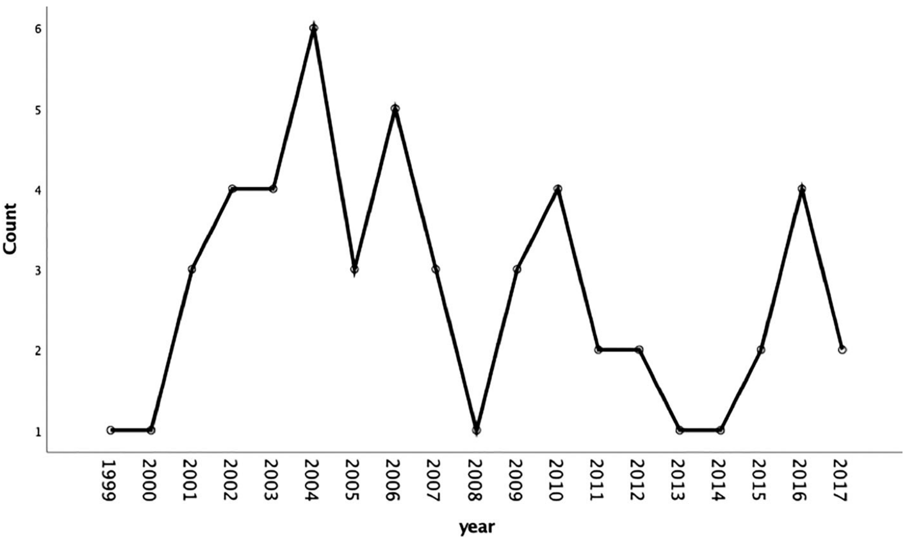

Fig. 1 Frequency of evaluation studies by year

and PricewaterhouseCoopers produced four studies, one of which was an IA. These two consultancies produced 27% of the total number of evaluation studies and four out of six of the consultancy reports commissioned for IAs.

Turning to systematic content analysis, 52 evaluation studies confirm that there are three different and distinctive types of documents. Since performance benchmarking is the main function of valuing-oriented evaluation studies (see the table in the online appendix), within this type of document graphs rank EU member states on their performance. For instance, the first evaluation study conducted in 1999 ranked Central and Eastern European countries' railway companies according to their productivity performance (study no. 1 in the online appendix). The EU-10 member states were ranked on a four-point scale of efficiency on several dimensions of labour and capital productivity. Twelve out of 15 valuing-oriented studies relied on visualisation (through graphs or tables) of comparative scoring of countries or railway companies' performance. Rather than cross-sectional data on policy performance, the remaining three studies are classified as valuing because they rely on the consultant's evaluation to identify best practices (study no. 35), they use performance measures for assessing the establishment of the European Railway Agency (study no. 39), and they are based on economic and financial analysis of investments for enhancing the clearance gauge of railways (study no. 51).

Use-oriented evaluation studies focus on stakeholders' use of the evaluation. This type of evaluation study tends to present graphs that are intended to provide an idea of the information process required to ensure the use of the proposed recommendation by stakeholders. For instance, the 2001 evaluation study conducted by Stratego (study no. 3) has the goal of keeping the railway companies' information burden to a minimum and presents a graph depicting a process of information to be provided by rail operators for establishing a pan-European railway information exchange sys-

Table 2 Frequency of evaluation study branches and chi-square test

Branch of evaluation	Observed frequency	Percentage	Expected frequency	Residual
Method	12	23.1	17.3	–5.3
Valuing	15	28.8	17.3	–2.3
Use	25	48.1	17.3	7.7
Total	52	–	–	–
Chi-square test				
	Value	Degrees of freedom	Significance	
Pearson chi-square	5.346	2	0.069	

tem. All of these use-oriented studies have a section on stakeholder consultation that is also common in many method-oriented and value-oriented studies. But what is distinctive in the former category of study is that the use of stakeholder consultation is the fundamental element of evaluation. Looking again at the study by Stratego, because stakeholders' goals are conflicting, this evaluation emphasises the acceptability of the proposed solutions by balancing stakeholders' views.

Finally, the content analysis of method-oriented evaluation studies confirms the general presence of formulas and models (such as cost and benefit analysis) for assessing the impact of a policy intervention. Ten out of 12 studies are based on a mathematical and statistical model. The two remaining studies instead revolve around a set of qualitative and quantitative impact analyses to support the creation of the European Railway Agency (study no. 41), and a classification of risk through risk acceptance criteria for the transport of dangerous goods (study no. 44). An example of this type of evaluation practice is a graph contained in a study conducted by Odegaard and Danneskiold-Samsoe (ODS) in 2002. This study (no. 7 in the online appendix) presents indicators of emitted noise that are calculated through a mathematical formula that discerns different types of vehicles and travel speeds. The final report and annex of this study are populated by technical specifications in order to set the technological state-of-the-art railway noise standards. Further, the report contains an attempt at assessing costs and benefits related to noise-reduction investments. This study fits with most features identified for method-oriented evaluation and EBPM. Overall, the evaluation theory tree has been tested on a medium-sized sample of evaluation studies composed of very diverse types of policy initiatives for addressing safety, market liberalisation and integration, and environmental issues.

Turning to descriptive statistics, Table 2 shows that use-oriented studies are the most frequent type of evaluation (25 studies representing 48% of the sample), followed by value-oriented studies (28.8%) and method-oriented studies (23.1%).

Accordingly, the EBPM wave of evaluation has not yet left much sediment on the shore of the European Commission's evaluation practice. The difference in distribution among the three types of evaluation studies is not statistically significant according to the conventional standard of $p < 0.05$. The one-sample chi-square value is equal to 5.346, with statistical significance at the level of $p = 0.069$. This result

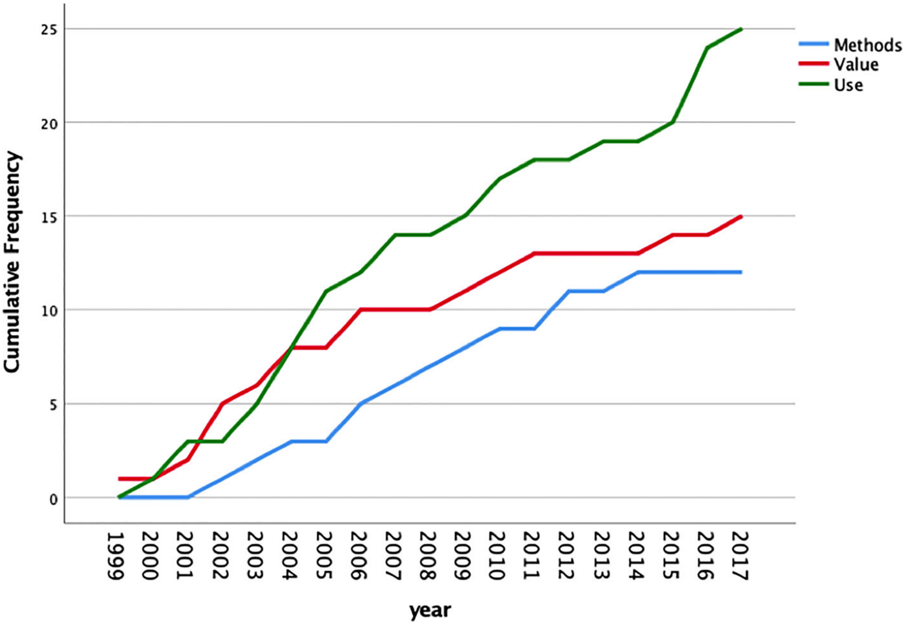

Fig. 2 Line graph of types of studies by evaluation theory branches and study year

shows that relying exclusively on the theoretical dimensions of evaluation is not sufficient to differentiate the distribution of the 52 evaluation studies, although the use-oriented studies are more frequent than the other two evaluation branches with 13 and 10 points of difference. The marginal statistical significance of this finding is also attested to by visually analysing changes in patterns in the type of evaluation studies used by the European Commission. Figure 2 shows that there is only a marginal change in the pattern of utilisation. The gaps between use-oriented studies and the other two types of studies tend to diverge in the most recent years. Looking at the trends of EBPM studies vis-à-vis NMP studies, one cannot detect any significant divergence across time.

Variety in the use of evaluation studies can be identified by focusing on policy sectors (Fig. 3) and their association with the three branches of evaluation in a contingency table (Table 3). Only two use-oriented studies could not be associated with one of the six sectorial types, lowering the number of observations for the contingency table to 50. The following qualitative observations can be derived from this analysis:

- Impact assessments are mainly conducted through method-oriented studies: 83.3%, in contrast to an average of 24%. In only one case did an external consultancy rely on benchmarking and value-oriented evaluation. In this small sample of evaluation studies, this finding confirms that BR is associated with EBPM rather than NPM.

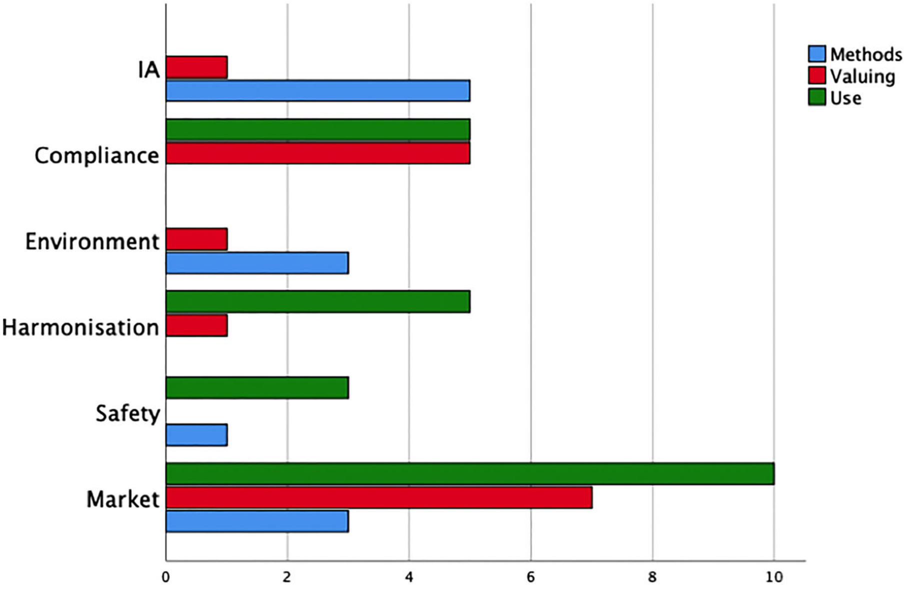

Fig. 3 Types of studies by policy sectors

Table 3 Cross-tabulation between evaluation theory branches and sectors of intervention and impact assessment

	Evaluation theory branches		
	Methods	Valuing	Use
Market	3 (15%)	7 (35%)	10 (50%)
Safety	1 (25%)	0	3 (75%)
Harmonisation	0	1 (16.7%)	5 (83.3%)
Environment and noise	3 (75%)	1 (25%)	0
Transposition and compliance	0	5 (50%)	5 (50%)
Impact assessment	5 (83.3%)	1 (16.7%)	0
Total	12 (24%)	15 (30%)	23 (46%)

- Use-oriented evaluation studies frequently address issues related to market, safety, and technical harmonisation, whereas no consultancy report on environmental issues relied on this theoretical and practical approach. It is important to note that the sample of evaluation studies for addressing environmental issues is small, and three out of the four total studies relied on the theory and practice of EBPM.
- Fifty percent of the transposition and compliance studies relied on the value-oriented approach, and the other 50% on use-oriented studies. Transposition and compliance are not analysed through method-oriented evaluation studies. This is a surprising finding since the large body of political science literature addresses the issue of (the lack of) compliance within the EU.
- Half of the market studies relied on use-oriented evaluation. This is a surprising finding as one would have expected more utilisation of the benchmarking method.

Table 4 Cross-tabulation for assessing the association between the new public management (NPM) and evidence-based policy making (EBPM) group and the group of market analyses and impact assessments (IAs)

	Use	NPM and EBPM	Total
Other analyses	13	11	24
	(54.2%)	(45.8%)	100.0%
Market analyses and IAs	10	16	26
	(38.5%)	(61.5%)	100.0%
Total	23	27	50
	46.0%	54.0%	100.0%
Chi-square tests			
	Value	Degrees of freedom	Significance
Pearson chi-square	1.239	1	0.266
Fisher's exact test	–	–	0.395

Given the small sample of observations and the presence of several cells with low and zero values, it is inappropriate to draw any conclusions about the association between evaluation branches and policy sectors despite a statistically significant (at the level $p < 0.05$) chi-square with a value of 28.3. To sort this methodological issue, the policy sectors have been recoded by collapsing market analyses and IAs into one group, with the remaining evaluation studies in the other group. The 2 × 2 contingency table (Table 4) now has frequencies over 5 for all cells, but the association between these groups is not statistically significant, as attested by the chi-square ($p = 0.266$) and Fisher's exact test ($p = 0.395$). Therefore, we can conclude that differentiation through the association between evaluation branches and policy sectors is mainly qualitative, as there are no patterns of association between EBPM and NPM evaluation on the one hand and market analyses and impact assessments on the other hand.

6 Discussion and Conclusion

This article has argued that documents such as evaluation studies are practised things that can provide useful information on patterns and trends of the EU evaluation. As demonstrated in the systematic content analysis of 52 consultancy reports on EU railway policy, evaluation practice can be distinguished according to three branches (methods, value, and use) of the evaluation theory tree. These theoretical branches have different functionalities and are associated with different diffusion waves of evaluation. In particular, this study has shown that there is an alignment between types and features (such as graphs, functions, and research design) of evaluation studies and communities of practice. Further, a specific type of evaluation reinforces a specific way of producing evidence and reinforcing epistemic knowledge.

The empirical findings of content analysis show that all three waves of evaluation are longitudinally well represented in the evaluation practice of consultants appointed by the European Commission since 1999. Although use-oriented evaluation is the most frequent type of study, the differences among the three categories are only marginally significant, and there are no notable patterns of association between evaluation branches and subtypes. However, the qualitative evidence of the contingency table shows that IAs mainly fall in the category of method-oriented studies. This finding contrasts with previous studies attesting the dominance of NPM for describing the EU IA system (Radaelli 2007, 2010). This difference in findings can be explained by the different level of analysis: Instead of focusing on the IA institutional system, this article has analysed IA documents of a specific EU policy vis-à-vis other types of evaluation conducted by external consultancies. The fact that most IAs analysed in this study are method-oriented does not mean, however, that they used the best scientific evidence. A previous study shows that the economic knowledge used in IAs on EU railway packages is not aligned with the scientific models for assessing the impact of liberalisation (De Francesco 2018).

Evaluation practices tend to remain stable across time, especially across the categories of NPM and EBPM. This finding contrasts with the expectation that evaluation diffusion waves transform the professional evaluators' methods and tools. Evaluation practice tends to rely on different functional purposes and professional prescriptive standards that tend to persist regardless of temporal and ideological contexts.

The proposed categorisation also enhances our conceptual understanding of the relationship among different elements of the EU evaluation system. For instance, the integration of ex ante appraisal and ex post evaluation can be achieved only through the evaluation practice associated with scientific and professional evidence. Indeed, the quality of (scientific) evidence and evaluation practice should be assured through a process of scientific peer review that also includes practitioners' and stakeholders' assessments. The quality of evidence and evaluation practice can be ensured by considering the experience of practitioners and of stakeholders' perceptions of policy effectiveness.

Overall, the evaluation theory tree has been tested on a small sample of evaluation studies composed of very diverse types of policy initiatives for addressing safety, market liberalisation and integration, and environmental issues. A comparative analysis of several economic and social policy sectors would allow us to overcome the main limitation of this study, i.e., its limited generalisation of the empirical findings. The use of the evaluation theory tree calls for additional research on other policy sectors and other type of documents, such as IAs conducted by the European Commission. Once categorised, IAs can be associated with the Regulatory Scrutiny Board's activities in overseeing and monitoring their quality.

References

Adler, Emanuel, and Vincent Pouliot. 2011. International practices. *International Theory* 3(1):1–36.

Alemanno, Alberto. 2015. How much better is better regulation. *Journal of European Risk Regulation* 6(3):344–356.

Arnold, Erik, John Clark, and Alessandro Muscio. 2005. What the evaluation record tells us about European Union framework programme performance. *Science and Public Policy* 32(5):385–397.

Bachtler, John, and Colin Wren. 2006. Evaluation of European Union cohesion policy: research questions and policy challenges. *Regional Studies* 40(02):143–153.

Bicchi, Federica. 2011. The EU as a community of practice: foreign policy communications in the COREU network. *Journal of European Public Policy* 18(8):1115–1132.

Bowen, Glenn A. 2009. Document analysis as a qualitative research method. *Qualitative Research Journal* 9(2):27–40.

Cairney, Paul. 2016. *The politics of evidence-based policy making*. London: Palgrave Pivot.

Christie, Christina A., and Marvin C. Alkin. 2008. Evaluation theory tree re-examined. *Studies in Educational Evaluation* 34(3):131–135.

Christie, Christina A., and Marvin C. Alkin. 2013. *Evaluation roots: a wider perspective of theorists? Views and influences*, 2nd edn., Thousand Oaks: SAGE.

Corradi, Gessica, Silvia Gherardi, and Luca Verzelloni. 2010. Through the practice lens: where is the bandwagon of practice-based studies heading? *Management Learning* 41(3):265–283.

Cunningham, P. 1997. The evaluation of European programmes and the future of Scientometrics. *Scientometrics* 38(1):71–85.

Dahler-Larsen, Peter. 2011. *The evaluation society*. Redwood City: Stanford University Press.

De Francesco, Fabrizio. 2018. Évaluer équitablement les régulateurs : aligner les analyses d’impact sur le savoir scientifique. *Politiques et Management Public* 35(3–4):131–151.

Desmarais, Bruce A., and John A. Hird. 2014. Public policy’s bibliography: The use of research in US regulatory impact analyses. *Regulation & Governance* 8(4):497–510.

Di Pietrantonio, Loris, and Jacques Pelkmans. 2004. The economics of EU railway reform. *Journal of Network Industries* 5(3–4):295–346.

Dunlop, Claire A., Martino Maggetti, Claudio M. Radaelli, and Duncan Russel. 2012. The many uses of regulatory impact assessment: a meta-analysis of EU and UK cases. *Regulation & Governance* 6(1):23–45.

Dyrhauge, Helene. 2013. *EU railway policy-making: on track?* Houndmills: Palgrave Macmillan.

Eversole, Robyn. 2012. Remaking participation: challenges for community development practice. *Community Development Journal* 47(1):29–41.

Finger, Matthias, and Pierre Messulam. 2015. *Rail economics, policy and regulation in Europe*. Cheltenham: Edward Elgar.

Freeman, Richard. 2009. What is ‘translation’? *Evidence & Policy* 5(4):429–447.

Freeman, Richard, and Jo Maybin. 2011. Documents, practices and policy. *Evidence & Policy* 7(2):155–170.

Freeman, Richard, Steven Griggs, and Annette Boaz. 2011. The practice of policy making. *Evidence & Policy* 7(2):127–136.

Head, Brian W. 2008. Three lenses of evidence-based policy. *The Australian Journal of Public Administration* 67(1):1–11.

Hoerner, Julian, and Paul Stephenson. 2012. Theoretical perspectives on approaches to policy evaluation in the EU: The case of cohesion policy. *Public Administration* 90(3):699–715.

Højlund, Steven. 2014. Evaluation use in evaluation systems—the case of the European Commission. *Evaluation* 20(4):428–446.

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

Huitema, Dave, Andrew Jordan, Eric Massey, Tim Rayner, Harro van Asselt, Constanze Haug, Roger Hildingsson, Suvi Monni, and Johannes Stripple. 2011. The evaluation of climate policy: theory and emerging practice in Europe. *Policy Sciences* 44(2):179–198.

Johansson, Kerstin, Verner Denvall, and Evert Vedung. 2015. After the NPM wave: evidence-based practice and the vanishing client. *Offentlig Förvaltning* 19(2):69–88.

Kallemeyn, Leanne M., Jori Hall, Nanna Friche, and Clifton McReynolds. 2015. Cross-continental reflections on evaluation practice: methods, use, and valuing. *American Journal of Evaluation* 36(3):339–357.

King, Jean A., and Laurie Stevahn. 2013. *Interactive evaluation practice: Mastering the interpersonal dynamics of program evaluation*. Thousand Oaks: SAGE.

Kirkpatrick, Colin, and David Parker. 2007. *Regulatory impact assessment: towards better regulation?* Cheltenham: Edward Elgar.

Lapsley, Irvine. 2009. New public management: the cruelest invention of the human spirit? *Abacus* 45(1):1–21.

Lee, Norman, and Colin Kirkpatrick. 2006. Evidence-based policy-making in europe: an evaluation of European commission integrated impact assessments. *Impact Assessment and Project Appraisal* 24(1):23–33.

Liberatore, Angela, and Silvio Funtowicz. 2003. ‘Democratising’ expertise, ‘expertising’ democracy: What does this mean, and why bother? *Science and Public Policy* 30(3):146–150.

Lægreid, Per, and Koen Verhoest. 2019. Reform waves and the structure of government. In *Public administration in europe: the contribution of EGPA*, ed. Edoardo Ongaro, 167–180. Cham: Palgrave MacMillan.

McGrath, Brian. 2016. Reflecting on ‘evidence’ and documentation devices in ‘translating’ community interventions. *Community Development Journal* 51(2):179–194.

Meads, Richard, and Lorenzo Allio. 2015. Paving the way to an improved, modern management of risk: the new European commission’s better regulation strategy. *European Journal of Risk Regulation* 6(4):649–651.

Mendez, Carlos, and John Bachtler. 2011. Administrative reform and unintended consequences: an assessment of the EU Cohesion policy ‘audit explosion’. *Journal of European Public Policy* 18(5):746–765.

Nash, Chris. 2008. Passenger railway reform in the last 20 years—European experience reconsidered. *Research in Transportation Economics* 22(1):61–70.

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. Edoardo Ongaro, Sandra Van Thiel, 577–593. London: Palgrave Macmillan.

Pollitt, Christopher. 2015. Wickedness will not wait: climate change and public management research. *Public Money & Management* 35(3):181–186.

Prior, Lindsay. 2008. Repositioning documents in social research. *Sociology* 42(5):821–836.

Radaelli, Claudio M. 2007. Whither better regulation for the lisbon agenda? *Journal of European Public Policy* 14(2):190–207.

Radaelli, Claudio M. 2010. Rationality, power, management and symbols: four images of regulatory impact assessment. *Scandinavian Political Studies* 33(2):164–188.

Radaelli, Claudio M. 2018. Halfway through the better regulation strategy of the Juncker commission: what does the evidence say? *Journal of Common Market Studies* 56(1):85–95.

Radaelli, Claudio M., and Fabrizio De Francesco. 2007. *Regulatory quality in europe: concepts, measures and policy processes*. Manchester: Manchester University Press.

Radaelli, Claudio M., and Anne Meuwese. 2012. How the regulatory state differs. The constitutional dimensions of rulemaking in the European Union and the United States. *Rivista italiana di scienza politica* 42(2):177–196.

Radaelli, Claudio M., Claire A. Dunlop, and Oliver Fritsch. 2013. Narrating impact assessment in the European Union. *European Political Science* 12(4):500–521.

Ryan, Katherine E. 2004. Serving public interests in educational accountability: alternative approaches to democratic evaluation. *American Journal of Evaluation* 25(4):443–460.

Sanderson, Ian. 2002. Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1):1–22.

Schoenefeld, Jonas, and Andrew Jordan. 2017. Governing policy evaluation? Towards a new typology. *Evaluation* 23(3):274–293.

Schoenefeld, Jonas, and Andrew Jordan. 2019. Environmental policy evaluation in the EU: between learning, accountability, and political opportunities? *Environmental Politics* 28(2):365–384.

Schrefler, Lorna, and Jacques Pelkmans. 2014. Better use of science for better EU regulation. *Journal of European Risk Regulation* 5(3):314–323.

Schwandt, Thomas A. 1997. Evaluation as practical hermeneutics. *Evaluation* 3(1):69–83.

Schwandt, Thomas A. 2005. The centrality of practice to evaluation. *American Journal of Evaluation* 26(1):95–105.

Shadish, William R., Thomas D. Cook, and Laura C. Leviton. 1991. *Foundations of program evaluation: theories of practice*. Newbury Park: SAGE.

Simons, Arno. 2016. Documented authority. The discursive construction of emissions trading in the expert literature. Doctoral dissertation. Berlin: Technische Universität Berlin, Berlin. https://depositonce.tu-berlin.de/bitstream/11303/5974/4/simons_arno.pdf. Accessed 21 Dec 2017.

Smismans, Stijn. 2015a. Opening editorial. *European Journal of Risk Regulation* 6(1):3–5.

Smismans, Stijn. 2015b. Policy evaluation in the EU: the challenges of linking ex Ante and ex post appraisal. *European Journal of Risk Regulation* 6(1):6–26.

Stame, Nicoletta. 2008. The European project, federalism and evaluation. *Evaluation* 14(2):117–140.

Strang, David, and John W. Meyer. 1993. Institutional conditions for diffusion. *Theory and Society* 22(4):487–511.

Torriti, Jacopo. 2010. Impact assessment and the liberalization of the EU energy markets: evidence-based policy-making or policy-based evidence-making? *Journal of Common Market Studies* 48(4):1065–1081.

Vedung, Evert. 2010. Four waves of evaluation diffusion. *Evaluation* 16(3):263–277.

Zwaan, Pieter, Stijn van Voorst, and Ellen Mastenbroek. 2016. Ex post legislative evaluation in the European Union: questioning the usage of evaluations as instruments for accountability. *International Review of Administrative Sciences* 82(4):674–693.