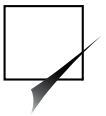

Article

Evaluation
2017, Vol. 23(1) 24–41

DOI: 10.1177/1356389016680877
journals.sagepub.com/home/evi

Evaluation capacity in the European Commission

Stijn van Voorst
Tilburg University, The Netherlands

Abstract
Ex-post evaluations are a potential tool to improve regulatory interventions and to hold rule-makers accountable. For these reasons the European Commission has promised to systematically evaluate its legislation, but it remains unclear if actual evaluation capacity is being built up in the Commission's Directorates-General. This article describes and explains the variation in evaluation capacity between the Directorates-Generals by applying a theoretical model of evaluation capacity developed by Nielsen et al. to the European context. To gain an in-depth understanding of the Directorates-Generals' evaluation capacity, 20 Commission officials were interviewed. The results show that there is much variation in the extent to which Directorates-Generals prioritize evaluation as well as in the amount of human and technological capital that they invest in evaluation. Further analysis using fuzzy-set Qualitative Comparative Analysis reveals that part of this variation can be explained by the Directorates-Generals' total budgets, suggesting that Directorates-Generals with a tradition of evaluating spending programmes also attach more importance to legislative evaluations.

Keywords
evaluation capacity, European Commission, European Union, fuzzy-set Qualitative Comparative Analysis (fsQCA), legislation, legislative evaluation

Introduction

Ex-post evaluations are a potential tool for improving legislation, as they can be used to learn about the implementation and the actual impact of regulatory interventions (Fitzpatrick, 2012: 480). Evaluations can also help to hold regulators accountable for their actions (Summa and Toulemonde, 2002: 409) and to control those who implement legislation (Stern, 2009: 76). These purposes of evaluation are especially relevant for the European Union (EU), which has

Corresponding author:
Stijn van Voorst, Tilburg University, Tilburg Law School, PO Box 91503, Tilburg, 5000LE, Netherlands.
Email: s.vanvoorst@fm.ru.nl

limited access to financial and communicative instruments and therefore relies heavily on legislative policies (Fitzpatrick, 2012: 489; Majone, 1999: 1).

Legislative evaluation is of particular importance to the European Commission, which bears the main responsibility for evaluation in the EU (Stern, 2009: 71). As an unelected body, the Commission has the constant need to show the added value of its policies (Scharpf, 1999: 187), for which ex-post evaluations can be a useful tool (Mastenbroek et al., 2016). Therefore, it is not surprising that the Commission has repeatedly stepped up its rhetoric in the field of ex-post legislative evaluations (Fitzpatrick, 2012: 489; Højlund, 2015: 40–4). In 2007 it pledged to evaluate not only its spending programmes, the evaluation of which has been common practice since the 1980s, but also non-spending activities like legislation (EC, 2007: 4; Højlund, 2015: 44). More recently the Commission (EC, 2013: 7; 2015: 17) has even promised to conduct evaluations of entire regulatory frameworks. However, existing research shows that these high ambitions may not always be realized. Available figures from the Commission (2013: 3) and academic research (Mastenbroek et al., 2016) indicate that the Commission has only evaluated about a third of the legislation that it should evaluate. Moreover, the quality of these evaluations seems to vary (Mastenbroek et al., 2016).

Theoretically, one explanation for variation in the initiation of evaluations lies in variation in evaluation capacity (Mastenbroek et al., 2016; Pattyn, 2014: 348). Evaluation capacity can be defined loosely as the presence of sufficient means and procedures for ensuring that evaluation and its appropriate uses are ordinary and ongoing (Stockdill et al., 2002: 14). It is a topic that has been studied in various settings, including non-profit organizations (e.g. Carman and Fredericks, 2010; Taylor-Ritzler et al., 2013), local and national governments (e.g. Bourgeois and Cousins, 2013; Nielsen et al., 2011) and international organizations (e.g. Taut, 2007). However, aside from a few sections in texts about the evaluation system of the EU (Stern, 2009: 71–2, 79–82; Summa and Toulemonde, 2002: 420–2; Toulemonde et al., 2005, 77–9) and evaluation use in the EU (Borrás and Højlund, 2015: 111; Technopolis, 2005: 45), there is little literature about the European Commission's capacity to evaluate, in particular when legislative evaluations are concerned.

Information provided by official sources is equally scarce. In 2007, the Commission reported that it had 140 full-time equivalents (fte) working on evaluation, with a total budget of €45 million (EC, 2007: 17), and that these numbers were gradually increasing because of investments made in trainings and networks (2007: 15–18). However, from this time onward the Commission has presented no more systematic data on its evaluation capacity, and it has never presented data on its capacity for legislative evaluations specifically.

This article seeks to fill this gap in our knowledge by answering the following questions: (1) to what extent do the Commission's Directorates-General (DGs) vary in their evaluation capacity? and (2) how can this potential variance be explained? The focus on DGs stems from the fact that they bear the main responsibility for conducting and outsourcing evaluations in the Commission (Stern, 2009: 71). The Commission's guidelines for evaluation – which were first adopted in 2004 and updated by the Commission's Secretariat-General (SG) in 2015 – require each DG to maintain an evaluation function with sufficient financial and human capital (EC, 2015: 268–88). However, in the end it is up to the DGs how they fulfil these requirements (Stern, 2009: 71). Therefore, variation between the Commission's DGs is crucial for understanding the functioning of legislative evaluation in the EU.

Building on an existing model (Nielsen et al., 2011: 326–7), this article splits evaluation capacity between evaluation demand and evaluation supply. Both concepts are quantified on

a scale of one to 50 points to allow for comparisons between DGs and to make the results useful for future research. Data were collected through in-depth interviews with 20 evaluation-related officials from 17 DGs responsible for legislation. The results not only provide a complete overview for 2014, but also show how evaluation capacity in the DGs has developed since 2000. The findings show that there is much variation between DGs in both the way in which they organize their evaluation-related procedures and the means which they invest in evaluation. Further analysis using fuzzy-set Qualitative Comparative Analysis (fsQCA) reveals that part of this variation can be explained by differences in the budgets of DGs, suggesting that DGs with a strong tradition in the field of evaluating spending programmes also have more capacity for legislative evaluation.

Theoretical framework

Selection of a model. A commonly used definition of evaluation capacity is that it is 'a system of guided processes and practices for ensuring that evaluation and its appropriate uses are ordinary and ongoing' (Stockdill et al., 2002: 14). Beyond such general definitions, however, evaluation capacity is a very ambiguous concept, with frequent debates among authors about what its main components are and how they should be measured (Nielsen et al., 2011: 324; Taylor-Ritzler et al., 2013: 192). In this article, the model of Nielsen et al. (2011: 328) will be used to measure evaluation capacity, as it has four key advantages which are relevant in the context of the European Commission. First, the model is meant to measure evaluation capacity at the organizational level (2011: 326). Since the aim of this article is to measure variation among the DGs of the Commission – organizational entities with their own evaluation policies – this focus on organizational aspects makes the model suitable for the research question at hand. Second, the model of Nielsen et al. (2011: 330) was created in the context of public sector organizations. Many other models published in recent years (e.g. Bourgeois and Cousins, 2013; Taylor-Ritzler et al., 2013) focus on non-profit organizations in the US, which only conduct programme evaluations and are therefore hard to compare to the EU. Third, the model allows evaluation capacity to be quantified on a scale of one to one hundred points, making it easy to compare capacity between organizations and making the results useful for future research. Fourth, the validity of the indictors of the model of Nielsen et al. (2011: 334–7) was thoroughly tested using factor analysis.

Besides these four advantages, there are three potential drawbacks to Nielsen et al.'s model. First, as the model is focused on organizational aspects, it ignores aspects of capacity related to single evaluations, such as the value that individual evaluation managers attach to learning (Bourgeois and Cousins, 2013: 299; Taylor-Ritzler et al., 2013: 192). However, because more than 200 ex-post legislative evaluations have been conducted in the EU between 2000 and 2012 alone (Mastenbroek et al., 2016), it would be impossible to measure indicators for every single evaluation. Second, the model is mostly focused on the minimum requirements that must be in place for evaluations to be embedded in an organization. Even when all these requirements are met, there is no guarantee that this will result in sound evaluations being produced and put to use. This should be remembered when interpreting the results. Third, the model was developed in the context of Danish local governments, meaning that it cannot be applied entirely to the EU level. In particular, the fact that most of the Commission's evaluations are outsourced (Stern, 2009: 69) had to be accounted for, resulting in some adaptions to the operationalization of the model which are further described below. However, most aspects

of the Nielsen et al.'s model could still be applied to the Commission, as its evaluation staff is explicitly required to be able to conduct internal evaluations and to scrutinize external evaluators whenever needed (EC, 2015: 268, 288).

Describing the model. Following other authors (e.g. Boyle et al., 1999: 11), Nielsen et al. distinguish between evaluation demand and evaluation supply (2011: 327). Evaluation demand refers to the fact that an organization considers evaluations valuable, while evaluation supply refers to the presence of sufficient means to evaluate (Nielsen et al., 2011: 326–7; Summa and Toulemonde, 2002: 422–3). In this context, 'means' refer to the staff responsible for evaluation and its methodological tools. Supply-side and demand-side conditions are equally important in determining evaluation capacity, as evaluations only come into existence when both interact, and therefore both are awarded 50 points in the model (Nielsen et al., 2011: 330).

Evaluation demand consists of two dimensions: the extent to which an organization has the explicit aim to evaluate (**evaluation goals**) and the extent to which evaluation is embedded in the daily functioning of an organization (**structure and processes**) (Nielsen et al., 2011: 326–7). Both dimensions are considered equally important in determining demand and are therefore awarded exactly 25 points (Nielsen et al., 2011: 330).

Evaluation goals, the first dimension of evaluation demand, consists of three main aspects (Nielsen et al., 2011: 328). The first is the amount of *formalization*: do official documents describe if and when evaluations must be conducted? A second aspect is the *utilization* of results, as evaluations are ultimately meant to be used, at least officially. Finally, the number of *evaluation purposes* stated by an organization is relevant. After all, evaluations can be used not only to create policies, but also to allocate resources, improve accountability and set priorities for the future (Nielsen et al., 2011: 180–4).

Structure and processes, the second dimension of evaluation demand, can be split into two aspects: the presence of an independent *evaluation unit* and the amount of *financial priority* that an organization attaches to evaluation (Nielsen et al., 2011: 330). Nielsen et al. also include the number of functions which an evaluator performs besides his core task in this dimension, but this aspect is left out in this article because in the EU most evaluations are conducted by external consultants.

Evaluation supply, the second condition of evaluation capacity, consists of two dimensions: the skills of those performing evaluations within an organization (**human capital**) and the non-human tools that allow evaluations to be performed (**evaluation technology**) (Nielsen et al., 2011: 327). Human capital (35 points) is more important than technology (15 points), as non-human tools are ultimately useless if there are no people who can apply them properly (Nielsen et al., 2011: 330).

When it comes to **human capital**, the first dimension of evaluation supply, three aspects are important: the *number of full-time employees* working on evaluations, the *evaluation trainings* completed by these employees and their *evaluation-related expertise* (Nielsen et al., 2011: 330). Nielsen et al. also include the formal education level of an organization's employees in this dimension, but this aspect is left out here as all the Commission's staff should have a master's degree, meaning there is little variation.

Evaluation technology, the second dimension of evaluation supply, consists of two aspects: the number of different *evaluation methods* (e.g. interviews, questionnaires) that are used and the application of any explicit *evaluation models* by an organization (Nielsen et al., 2011: 327; for examples of models, see Fitzpatrick, 2012: 481). Nielsen et al. also include the

Table 1. Model of evaluation capacity (p = points).

Condition	Dimension	Aspect
Evaluation demand	Evaluation goals (25p)	Evaluation purposes (8p)
		Formalization (7p)
		Utilization (10p)
	Structure and processes (25p)	Evaluation unit (9p)
		Financial priority (16p)
Evaluation supply	Human capital (35p)	Number of full-time employees (10p)
		Evaluation training (10p)
	Evaluation technology (15p)	Evaluation-related expertise (15p)
		Evaluation methods (10p)

presence of evaluation software in this dimension, but since EU evaluations are usually outsourced this aspect was irrelevant for this study. Table 1 summarizes the model as it was used in this article.

Explaining evaluation capacity. Aside from describing the variation in evaluation capacity between the Commission's DGs, this article also seeks to explain it. Although there is no single theoretical framework for this purpose, three separate explanations can be derived from the literature: the amount of legislation to be evaluated (functionalist logic), the presence of a tradition of evaluating spending programmes (historical institutionalism) and the sensitivity of the DG's policies (political rationality).

First, following a functionalist logic, the *amount of legislation* that has to be evaluated by a DG could influence its evaluation capacity. Since evaluation is compulsory for most EU legislation (EC, 2015: 261), it can be expected that DGs responsible for more legislation will have more evaluation demand (Hypothesis 1) and supply (Hypothesis 2).

Second, the extent to which an organization has a *tradition of evaluating spending programmes* is a possible explanation for variation in evaluation capacity, since building such capacity is a long-term investment (Preskill and Boyle, 2008: 451). Because EU evaluations have their origins in the field of spending programmes, the DGs that spend most money have the longest experience with evaluation (Fitzpatrick, 2012: 479; Stern, 2009: 69). Therefore, it is expected that DGs with a stronger tradition of evaluating spending programmes also have more demand (Hypothesis 3) and supply (Hypothesis 4) for legislative evaluations. This argument follows the logic of historical institutionalism: the policies of an organization are bound by the decisions which it made in the past (Nugent, 2010: 438).

Third, evaluation capacity could be influenced by the *political sensitivity of the DG's policy area*. The Commission is often assumed to follow a strategy of legislative expansion: because it lacks strong financial or communicative instruments, it tends to focus on expanding EU law to encourage European integration (Majone, 1999: 65). Evaluations are a potential threat to this strategy, as they can be used as an argument to roll back policies in case their findings are negative (Weiss, 1993: 94). This idea is closely linked to the political rationality of evaluations: evaluations are not neutral objects, but can be used to threaten or defend the interests of actors in the policy process (Weiss, 1993: 94; Bovens et al., 2008: 320). Bureaucracies may try to avoid them when they threaten to undo the results of previous political investments or

negotiations (Weiss, 1993: 95–6). Following this logic, it can be expected that DGs dealing with sensitive policies areas will have less evaluation demand (Hypothesis 5) and supply (Hypothesis 6).

Method

Data collection

Empirical data were gathered through face-to-face interviews with the main coordinator of ex-post legislative evaluations in each DG (17 in total). Although it was attempted to also speak with the head of the evaluation function of each DG, this was only possible in three cases: most heads of unit referred back to their coordinator for legislative evaluation because the requested information was highly detailed. Interviews with three DGs could only be conducted by phone or email.[1] The data provided by the respondents were always checked by using available documentation about the evaluation policies of the DGs (such as guidelines and annual activity reports). Such documents were usually found on the DGs' websites, but respondents were also asked to provide additional documents. In three cases, an indicator could be measured only through online documents.[2]

Since this article focuses on legislative evaluations, only DGs responsible for major legislation were included. To find out which DGs meet this requirement, a self-constructed dataset of European regulations and directives from 2000–14 was used (Mastenbroek et al., 2016). The dataset excludes amendments, rectifications, implementing legislation, repeals, and legislation concerning individual countries, because such small acts are rarely evaluated (Stern, 2009: 71). Using the online database Eurlex, each piece of legislation was linked to the DG which initiated it. Only DGs appearing at least once in this way were included in the research. DGs dealing with foreign affairs or the Commission's internal functioning were excluded, as their legislation is not aimed at citizens and therefore follows a different logic concerning evaluation. Applying these criteria, the study focusses on 17 DGs existing in 2014: Agriculture (AGRI), Communications and Technology (CONNECT), Competition (COMP), Economic and Financial Affairs (ECFIN), Employment, Social Affairs and Inclusion (EMPL), Energy (ENER), Enterprise and Industry (ENTR), Environment (ENV), Eurostat (ESTAT), Home Affairs (HOME), Justice (JUST), Maritime Affairs (MARE), Internal Market (MARKT), Mobility and Transport (MOVE), Health and Consumers (SANCO), Taxation (TAXUD) and Trade (TRADE).

Operationalization

To operationalize the aspects of evaluation capacity described in the theoretical framework section, the indicators used by Nielsen et al. (2011: 330–1) and their relative weights were used as much as possible (Ramboll Management Consulting, 2011: 2). However, four adaptions were made to fit the model to the specific context of this study. First, since the exact number of evaluations conducted by the Commission is unclear (Mastenbroek et al., 2016), all indicators requiring knowledge about numbers of evaluations were removed. Second, since evaluations in the EU are often outsourced (Stern, 2009: 71), two different indicators were used to measure evaluation-related expertise. Third, the indicators for evaluation-related trainings and utilization were dichotomized because some collected data for these indicators were unspecific. Fourth, the indicator for evaluation models was changed to specify what an

'evaluation model' means in the context of the EU. When an indicator was removed or added to the model, the number of points awarded to the other indicators inside the same dimension was increased or decreased proportionally.

The aspect of *formalization* was measured by asking if the DG has an official planning for future legislative evaluations (4 points) and any formal rule for when legislation should be evaluated when this is not compulsory (3 points). The aspect of *utilization* was measured by asking if there is a standardized procedure for employees of the DG to respond to results of legislative evaluations (10 points). *Evaluation purposes* were measured by asking the respondents what aims legislative evaluation has in their DG. Improving policies, increasing accountability, efficient resource allocation, political supervision, long-term learning and setting priorities all qualify as different purposes (8 points).

The presence of an independent *evaluation unit* was measured by asking if the DG has a unit or subunit for which ex-post evaluation and related issues (like ex-ante evaluation) are its core task (9 points). Although each DG must have an evaluation function (EC, 2007: 16), this does not necessarily take the form of a specialized unit, so there is room for variation here. *Financial priority* was mapped by asking how much money each DG spends on an average ex-post evaluation of one regulation or directive (16 points). There are no hard standards for this kind of expenditure in the EU, so the DG with the highest expenditure per evaluation was used as a benchmark and other scores were adapted proportionally.

The *number of fte* working on evaluation was measured by asking how many people (in fte) work for the centralized evaluation function of each DG (10 points). Although other employees can also spend time on evaluations, their work is too fragmented to measure. The aspect of *evaluation training* was measured by asking if the DG organizes any evaluation trainings (10 points). *Evaluation-related expertise* was measured by asking in how many evaluation-related networks the DG's employees participate, as such networks are an important way of building expertise (Stern, 2009: 71) (9 points). To measure the external expertise available to each DG, the average number of external companies that bid for its legislative evaluation was asked for (6 points).

The aspect of *evaluation methods* was measured by checking if the DG has guidelines on how to conduct ex-post legislative evaluations in its policy field, with 2 points awarded per method described (10 points). The number of methods is relevant here because Commission officials must be able to scrutinize a broad range of external evaluations (EC, 2015: 288). The aspect of *evaluation models* was measured by checking if the DG has any written guidelines for modelling causal effects in legislative evaluations (5 points). Since all legislative evaluations of the Commission are supposed to map causality, this indicator is relevant for each DG (EC, 2015: 53). For both evaluation methods and models, the number of points awarded is halved if the DG only has guidelines for ex-post evaluation in general. The operationalization is summarized in Table 2.

As for the explanatory conditions, the *amount of legislation* was measured via the number of major regulations and directives initiated by each DG over the period 2000–14. To measure this, the self-constructed dataset of legislation which was already described above was used. The extent to which a DG has *a tradition of evaluating spending programmes* was estimated through the size of its budget, which was retrieved from the annex of each DG's annual activity report.[3] The *political sensitivity of the DG's policy area* was measured by looking at the policy field which is handled by each DG. In the EU, some policy fields are (partly) dealt with through unanimity voting in the Council because they are related to the sovereignty of the

Table 2. Operationalization of evaluation capacity (p = points).
Indicators highlighted with an asterisk were measured specifically for legislative evaluations; other indicators could only be measured for ex-post evaluation in general.

Dimension	Aspect of capacity	Indicator measured during interviews (max 100p)
Evaluation goals (25p)	Evaluation purposes (8p)	Number of purposes of legislative evaluation stated (2p per purpose)*.
	Formalization (7p)	Presence of evaluation planning (4p, yes/no)*.
		Presence of guidelines about when to evaluate legislation (3p, yes/no)*.
	Utilization (10p)	Presence of a procedure for responding to legislative evaluation results (yes/no)*.
Structure and processes (25p)	Evaluation unit (9p)	Presence of a unit for which ex-post evaluation is a core task (yes/no).
	Financial priority (16p)	Money spent on an average evaluation of one regulation or directive as a % of the money spent by the DG with the highest expenditure*.
Human capital (35p)	Number of full-time employees (10p)	Number of fte working for centralized evaluation function (2p per fte).
	Evaluation training (10p)	Existence of a DG-level evaluation training (yes/no).
	Evaluation-related expertise (15p)	Number of evaluation-related networks (9p maximum, 3p per network).
		Number of external companies which bid for legislative evaluations of the DG (6p maximum, 1p per company)*.
Evaluation technology (15p)	Evaluation methods (10p)	Presence of internal guidelines on evaluation methods (10p, 2p per method)*.
	Evaluation models (5p)	Presence of internal guidelines on evaluation models present (yes/no)*.

member states, which means they are considered sensitive issues. This mostly concerns justice and home affairs, social policies and taxation, so the DGs dealing with these topics (JUST, HOME, EMPL and TAXUD) were coded as sensitive while the other DGs were coded as non-sensitive (Nugent, 2010: 308).

Method

Fuzzy-set QCA (fsQCA) was used to explain evaluation capacity,[4] for two reasons. First, this method is useful for mapping the various combinations of causal conditions that can explain variation in a given outcome (in this article: evaluation demand and supply). Second, while the sample of 17 DGs is too small to allow for regression analysis, it is large enough to make fsQCA a feasible method (Ragin, 2008: 9).

FsQCA allows for the use of continuous scales if they are transformed to vary between zero and one. Through the so-called direct method of transformation, conditions can be transformed if appropriate values are derived from the literature for the scores of zero (non-membership), 0.5 (cut-off point) and one (full membership) (Ragin, 2008: 85). Since the presence of political sensitivity is measured dichotomously in this article, the only conditions that require transformation are the presence of a large amount of legislation, the presence of a large budget, and the two outcomes (the presence of high evaluation demand and high evaluation supply).

DG Environment (ENV) is a prime example of a DG with a large amount of legislation. Because the aim of environmental policy is to regulate and prevent polluting behaviour from citizens and companies, the DG is responsible for a large number of regulations and directives in the fields of waste, water quality, chemicals, noise, genetic manipulation and biodiversity (Nugent, 2010: 346–50). Therefore, the observed number of legal acts of DG environment (76), the second highest in the data, was used as the score for full membership.

DG Agriculture (AGRI) is the classic example of a DG which relies heavily on spending. To support European farmers and make their activities sustainable, the DG uses a mix of direct payments, refund operations, rural development programmes and other subsidies. Together agricultural policies account for 40 per cent of the EU budget. Therefore, the observed budget of DG AGRI (about €58 billion in 2013) represents full membership (Nugent, 2010: 353–63).

DG Home Affairs is a prime example of a DG with a medium-sized amount of legislation and a medium-sized budget. The DG aims to halt crime and illegal migration, which requires numerous regulations about cooperation between member states – the Dublin regulation on migration being one example (50 observed legal acts in the data). However, the DG also manages financial activities such as the European refugee fund and the European return fund (observed budget: about €1 billion) (Nugent, 2010: 335–9). Therefore, the observed values of DG Home Affairs were used to represent the cut-off points for the explanatory conditions. This leaves four DGs which spend several billions in the group with 'high budgets' (the DGs for agriculture, technology, employment and enterprise), and a larger group of 11 DGs which spend 'just' a few hundred million euros in the group with 'low budgets'. The second group of DGs includes many DGs that rely on a large amount of legislation (70–90 acts) rather than high budgets for their policies (i.e. the DGs for environment, the internal market, health affairs and infrastructure).

There is also a group of seven DGs with neither a high budget nor a high amount of legislation. DG Competition is a prime example of this. Its main activities are decisions to allow or forbid state aid and company mergers, which requires only a handful of regulations (7) and a small budget (about €5.6 million in 2013) (Nugent, 2010: 327–8). Therefore, its observed values were used to represent non-membership for the explanatory conditions.

Evaluation demand and supply are transformed by dividing their scores by 50, which can be done for such self-constructed scales (Kogut et al., 2004: 123). Because all the cut-off points presented in this section are to some extent arbitrary – there is no strong set-theoretical knowledge for any of these conditions – alternative cut-off points will also be tested during the analysis to see how this affects the results.

Results

This section presents empirical results for the indicators for evaluation capacity listed above. For DG ECFIN, data on all indicators measured specifically for legislative evaluations are missing, as this DG will only start evaluating its legislation in the near future. For DG ESTAT there is no data on the number of external companies bidding for legislative evaluations and the amount of money spent on an average evaluation, as all its legislation is evaluated internally and no consultant is paid. Table 3 and 4 summarize the results.

Concerning *evaluation purposes*, all DGs mentioned the improvement of policies and the need to be accountable to the Council and the European Parliament as important aims of legislative evaluation. About half of all DGs also mentioned setting political priorities as a

Table 3. Results for evaluation demand (2014).

DG	Evaluation purposes	Planning	When evaluate	Utilization	Evaluation unit	Financial priority
AGRI	3	Yes	Yes	No	Yes (1998)	400.000
CONNECT	3	Yes	Yes	Yes	No	200.000
COMP	4	Yes	No	No	No	225.000
ECFIN	–	–	–	–	Yes (2005)	-
EMPL	2	Yes	No	Yes	Yes (1998)	300.000
ENER	2	Yes	No	Yes	No	250.000
ENTR	3	Yes	Yes	Yes	No	260.000[2]
ENV	3	Yes	Yes	No	No	250.000
ESTAT	2	Yes	No	Yes	Yes (2005)	-
HOME	3	Yes	No	No	No	350.000
JUST	2	Yes	Yes	No	No	200.000
MARE	2	Yes	No	No	No	200.000
MARKT	3	Yes	Yes	Yes	Yes (2008)	240.000[2]
MOVE	2	Yes	No	Yes	Yes (2007)	195.000[2]
SANCO	2	Yes	Yes	Yes	No	200.000
TAXUD	2	Yes	No	No	Yes (2010)	210.000
TRADE	3	Yes	No	Yes	No	175.000

Table 4. Results for evaluation supply (2014).

DG	Evaluation training	Fte	Number of networks	Number of bids	Method guidelines	Model guidelines
AGRI	No	15	3	5	None	No
CONNECT	Yes (2013)	2	3	10	10 (2011)	Yes (2011)
COMP	Yes (2013)	2	3	2	None	No
ECFIN	Yes (2008)	2	1	-	-	-
EMPL	Yes (2004)	4.5	3	4	6 (2001)	Yes (2001)
ENER	No	1	1	5	None	No
ENTR	No	2	2	1	12 (2002)	Yes (2002)
ENV	Yes (2007)	1.4	1	20	None	Yes (2003)
ESTAT	No	2	2	-	None	No
HOME	No	1.3	2	5	12 (2011)	Yes (2011)
JUST	Yes (2011)	0.25	1	5	None	Yes (2011)
MARE	No	0.5	2	1	None	No
MARKT	Yes (2008)	1	1	6	15 (2008)	Yes (2008)
MOVE	No	2.5	1	4	None	No
SANCO	No	2	1	3	None	None
TAXUD	Yes (2012)	2.5	1	3	None	No
TRADE	Yes (2011)	1	1	6	3 (2008)	No

purpose of legislative evaluation. Some DGs mentioned other aims as well, such as basic learning (COMP, TAXUD) and efficient resource allocation (AGRI).

As for *formalization*, all DGs have a planning for future legislative evaluations as a part of their annual management plans, as this practice is enforced by the SG. Some DGs publish

this part of their annual management plan online, while others keep it internal. Seven DGs have guidelines stating after how many years a piece of legislation should be evaluated, which was between five and seven years in all cases. For other DGs an evaluation is generally initiated only when an evaluation clause makes this compulsory or when revision appears necessary.

Concerning *utilization*, nine DGs have an official follow-up procedure which applies to legislative evaluations. This usually takes the form of a requirement to write an action plan or fiche by the main policy unit involved in the evaluation, although in DG SANCO the plan is sometimes written by the evaluation unit and in DG EMPL it is a joint responsibility. Such actions plans usually require approval at the management level. For DGs without a follow-up procedure, legislative evaluations are often followed directly by an impact assessment (an obligatory ex-ante evaluation of legislative amendments) or by no action at all.

Of all the DGs examined, only DG AGRI has a *unit fully dedicated to ex-post evaluation*. DG ECFIN, EMPL, ESTAT, MARKT, MOVE and TRADE have units responsible for ex-post evaluation and related issues (like impact assessments), while for other DGs the evaluation function is located together with broad support functions like finances or strategy. DG AGRI also leads when it comes to *financial priority*, reporting the highest budget for an average evaluation of one regulation or directive (€400.000) because it often requires case studies in each member state. DG HOME also reports a high budget per evaluation (€350.000), while the other DGs vary between €175.000 and €300.000. This data should be interpreted with some caution, as budgets show much variation.

As Table 4 shows, nine DGs organized an *evaluation training* in 2014. Most of these trainings were set up during the last few years. Other DGs only participate in the centralized training organized by the SG (five days total), which is compulsory for all evaluation-related staff. Because trainings at the DG-level usually last one day at most, they are more suitable to reach a broad audience of policy makers.

Concerning the *number of fte* working for the evaluation functions, it turned out that in some DGs coordinating evaluations in only a part of the job of a single person (MARE, JUST), while others dedicate a small team to the issue (EMPL, ENTR, TAXUD, MOVE). Staff differences are present in relative terms as well: DG Agriculture has about 1.4 per cent of its staff working on coordinating evaluations, while for DG JUST this is 0.05 per cent.[5] It should be noted, however, that the high numbers for DG AGRI and CONNECT are partly caused by the fact that their evaluation functions conduct some programme evaluations themselves, rather than only supporting other units.

Concerning *evaluation expertise*, all DGs participate in the central evaluation network organized by the SG, but beyond that there is much variation. DG COMP, HOME, MARE and EMPL have internal evaluation networks in which their various directorates are represented, while DG AGRI and EMPL organize networks with member state evaluation experts. DG COMP contributes to an OECD evaluation network, while DG ESTAT, ENTR and CONNECT participate in evaluation-related networks of all DGs working in a specific policy area. Other DGs only work with infrequent (lunch) sessions to discuss evaluation.

Concerning external expertise, most DGs work with framework contracts for their legislative evaluations, meaning that only a limited number of preselected companies (between three and six) may bid for contracts. DG CONNECT and DG ENV usually allow open competition between companies, which explains the high number of average bids they receive. Using open competition can take twice as much time as using framework contracts, which is why most

DGs prefer the latter, although most respondents do believe that going to the market offers extra quality.

As for *evaluation methods* and *evaluation models*, only DG CONNECT and DG MARKT have guidelines about these topics specifically for legislative evaluation. DG EMPL, ENTR, ENV, HOME, JUST, SANCO and TRADE have internal guidelines for ex-post evaluation in general. While DG MARKT, CONNECT, HOME and ENTR discuss 10 or more methods in their guidelines, the guidelines of other DGs discuss only a small number of methods or no methods at all, and the document of DG SANCO discusses neither methods nor models.

A question that remains is whether these results based on 2014 are still up-to-date, as the SG published new evaluation guidelines in May 2015 (EC, 2015). None of the respondents believed that the new guidelines would lead to immediate human or financial investments in evaluation at the DG-level. However, as the new guidelines do specify rules for writing follow-up action plans (EC, 2015: 297–8), the current variation among DGs on that aspect might be reduced. Furthermore, at the beginning of 2015 DG HOME and DG TRADE have created units dealing specifically with evaluation and related matters to reflect the growing importance of these topics.

Using the relative weights of the indicators listed in Table 2, Figure 1 provides the final scores for each DG on evaluation demand, evaluation supply and total evaluation capacity. DGs are ranked from highest to lowest, but the scores should be seen as descriptions rather than judgements. The results show a large group of DGs receiving between 45 and 55 points in total, with a few outliers having higher and lower scores. Values for evaluation demand are generally a little higher than for evaluation supply, especially for the DGs at the lower end of the spectrum, indicating that these DGs may deem evaluation important but have little means to invest in it.

Explanatory analysis

When applying fsQCA, a useful first step is to test if any individual explanatory conditions are either necessary or sufficient to let the outcome occur. Therefore, Table 5 shows which conditions provide consistent explanations for high evaluation demand and supply (benchmark: >0.80) and which consistency scores above the threshold are probabilistically significant (benchmark: <0.05). Unlike the consistency threshold, the probabilistic test also takes the number of cases in which a condition leads to an outcome into account (Ragin, 2008: 120), which makes it useful for this article because some of the conditions are only present in a handful of cases. Since in fsQCA the absence of an explanatory condition does not necessarily have the opposite consequences of the presence of that condition, the negation of each condition is also included in the table, recognizable by the symbol ~.

As the results show, neither the amount of legislation nor the political sensitivity of a DG nor their negations are necessary or sufficient conditions for high evaluation demand or supply (Hypotheses 1, 2, 5 and 6 are falsified). This outcome is also apparent if we look at the four DGs with the highest overall capacity: DG EMPL, CONNECT, MARKT and AGRI. In the first two of these DGs the condition of having a large amount of legislation is clearly absent (both DGs initiated about 25 regulations and directives, while the cut-off point is 50), and out of the four cases only DG EMPL deals with a policy field that is considered politically sensitive.

However, in line with Hypothesis 3, the presence of a high budget is a sufficient condition for high evaluation demand. In other words, when a DG has many financial resources we can

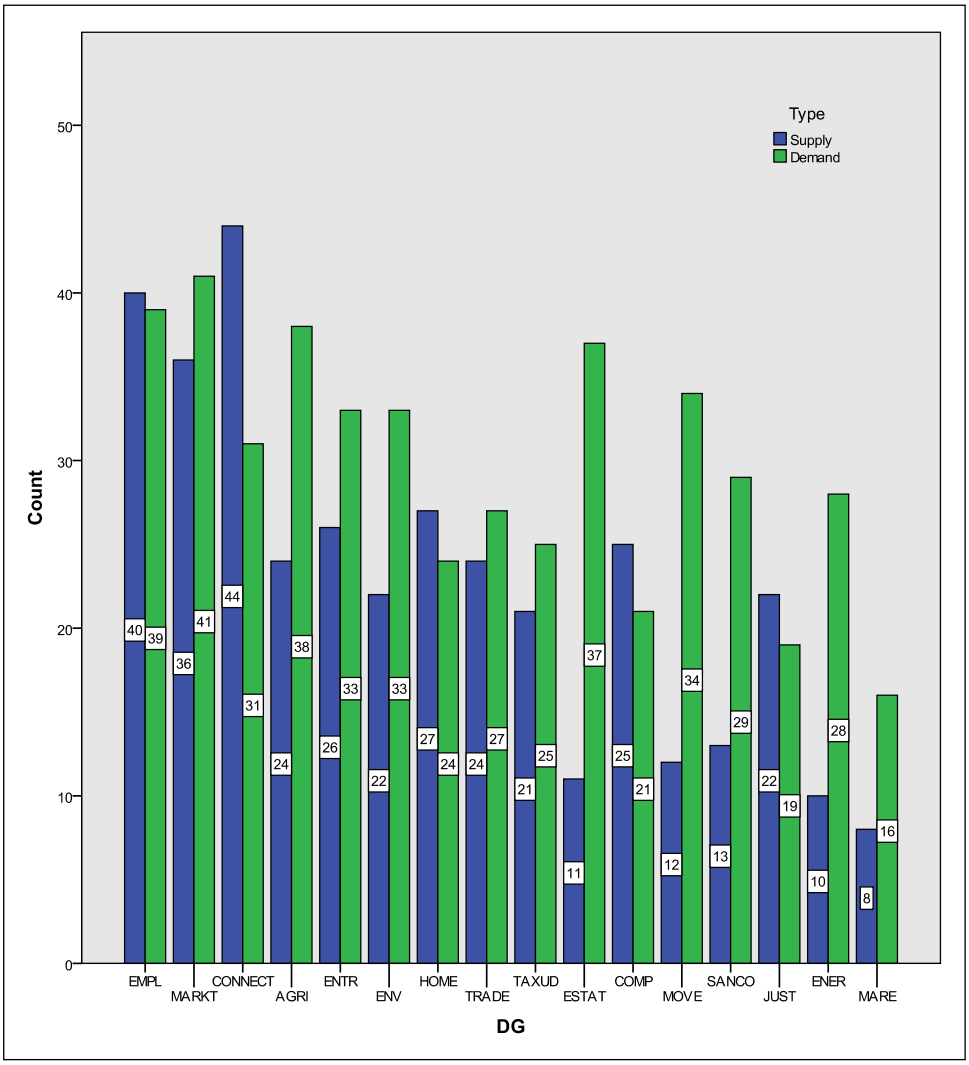

Figure 1. Scores on evaluation supply and demand for each DG.
Values for ESTAT are adapted proportionally to compensate for missing data.

Table 5. Results of QCA analysis for evaluation demand/evaluation supply.
The level of significance for all proportions >0.80 is provided in parenthesis. ~ represents the negation of a condition.

Condition	Consistency (necessary conditions) for evaluation demand/evaluation supply	Consistency (sufficient conditions) for evaluation demand/evaluation supply
High legislative amount	0.63/0.63	0.79/0.62
~High legislative amount	0.63/0.72	0.69/0.61
High budget	0.44/0.51	0.95 (0.001)*/0.84 (0.65)
~High budget	0.85 (0.56)/0.82 (0.85)	0.70/0.51
High sensitivity	0.23/0.30	0.54/0.55
~High sensitivity	0.78/0.70	0.61/0.43

Table 6. Truth table.

High amount of legislation	High budget	High sensitivity	N	Cases	Outcome (high demand)	Outcome (high supply)	Raw consistency (demand/supply)
No	No	No	5	COMP, ENER, ESTAT, MARE, TRADE	No	No	0.69/0.48
Yes	No	No	4	ENV, MARKT, MOVE, SANCO	Yes	No	0.83/0.60
Yes	Yes	No	2	AGRI, ENTR	Yes	Yes	1.00/0.95
No	No	Yes	2	JUST, TAXUD	No	No	0.72/0.72
No	Yes	Yes	1	EMPL	Yes	Yes	0.98/1.00
No	Yes	No	1	CONNECT	Yes	Yes	0.99/0.88

expect it to attach much importance to legislative evaluation. The four DGs with budgets of more than €1 billion (EMPL, CONNECT, ENTR and AGRI) all have high evaluation demand as well. The corresponding coverage score is 0.44, showing that the high budget condition accounts for a little less than half of the cases with high evaluation demand.

The presence of a high budget is also a sufficient condition for high evaluation supply, in line with Hypothesis 4. This relationship does not pass the probabilistic test, but when taking a closer look at the data that fact is explained entirely by the case of DG AGRI, which has a very high budget and only a medium score (24 points) on evaluation supply. Generally speaking, it therefore seems that DGs with a high budget also invest a large amount of human and technological capital in legislative evaluations. The corresponding coverage score is 0.51, showing that the high budget condition accounts for about half of the cases with high evaluation supply.

According to the theoretical framework of this article, these results indicate that DGs with a tradition of evaluating spending programmes also have high capacity for legislative evaluations. This interpretation is supported by statements made by several respondents during the interviews. For example, one evaluation coordinator from a low-budget DG stated that evaluation used to be a low priority in his organization because the evaluation culture in the Commission was so focused on accounting for how money was spent. Only in the last four years a shift began towards legislative evaluations, and since then his DG has slowly started building evaluation capacity. Another respondents emphasized that his DG already conducts evaluations since the 1990s, starting with spending programmes, and has therefore become a frontrunner concerning all kinds of ex-post evaluations in the Commission.

Besides looking at individual conditions, fsQCA also allows for analysing combinations of causal conditions (Ragin, 2008: 125). The truth table above (Table 6) shows all combinations that appear in the data with their corresponding cases. Only those combinations with a consistency score of more than 0.8 were included in the analysis, to ensure that the results remain undistorted by combinations with contradictory scores on evaluation demand and supply (2008: 139).

To analyse the truth table, the intermediary method was used (Ragin, 2008, 164).[6] The results show that the absence of political sensitivity in combination with the presence of a large amount of legislation consistently leads to high evaluation demand (consistency = 0.77) and that this solution covers about one-third of the cases with high demand (unique coverage = 0.29). This

result indicates that DGs will usually value legislative evaluation if they have a strong legislative responsibility in a policy field which is not so sensitive that evaluations might be threatening. However, more research is needed to confirm if this interpretation is correct.

When including the presence of high budgets, the entire solution for high evaluation demand has a consistency of 0.80 and a coverage of 0.73, meaning that it covers about three-quarters of the cases with high evaluation demand. No combinations of conditions were found which explain the presence of high evaluation supply.

To check the robustness of these findings, various higher and lower cut-off points for the explanatory conditions were tested.[7] The results were largely unaffected by these changes: the presence of a high budget kept being a sufficient condition for high demand (consistency >0.9), while most other individual conditions remained neither necessary nor sufficient. However, if the cut-off point for high budgets is put below €775 million, which is close to the medium-sized budget of a case like DG MARE, it seizes to be a sufficient condition for high evaluation supply.

The analysis so far focused on explaining the presence of high evaluation demand and supply, which in fsQCA is not the same as explaining the absence of these outcomes (Ragin, 2008, 102). A similar analysis was therefore conducted for the negations of the outcomes. Its results cannot be fully presented here due to word constraints, but it can be said that its results were in line with the previous findings. The absence of high budgets turned out to be a necessary condition for both low evaluation demand (consistency = 0.96; $\alpha = 0.00$) and low evaluation supply (consistency = 0.92; $\alpha = 0.05$), indicating that almost all DGs with low evaluation capacity also have low budgets (DG AGRI being the only exception). No other individual conditions consistently explained the absence of the outcomes, nor did any combinations of conditions.

Conclusion

This article addressed the questions how the DGs of the European Commission vary in their capacity for legislative evaluations and how this variance can be explained. Through in-depth interviews with 20 Commission officials, data were collected about both evaluation supply and demand. The results reveal much variance in capacity between DGs. On the demand side, some DGs have very clear aims and procedures for all the stages of legislative evaluation, while for other DGs this is not the case. On the supply side, while for some DGs coordinating evaluation is a part-time job of one person, others devote a small team or even a whole unit to the task. Over the last few years the number of DGs supporting their staff with their own evaluation-related trainings, networks and guidelines has gradually increased, but each of these features is still present in only about half of the DGs. The highest overall evaluation capacity for 2014 was found in DG EMPL, MARKT, CONNECT and AGRI.

How can this variance be explained? The analysis shows that the presence of a high budget is a sufficient condition for high evaluation demand and supply, indicating that DGs with a long tradition of evaluating spending programmes attach more importance to and invest more means in legislative evaluations than other DGs. Theoretically there could be other explanations for the relationship between high budgets and high capacity, but the qualitative information from the interviews confirms that DGs with a long tradition of evaluating spending activities also pay more attention to legislative evaluations today. Furthermore, the analysis indicates that DGs with have a strong legislative responsibility in a policy field which is not

very sensitive usually have high evaluation demand. However, more qualitative research should be conducted to verify this interpretation.

Two other possibilities for future research stand out. First, this article was mostly focused on the presence of certain minimal organizational requirements to systematize legislative evaluation in the Commission's DGs. It has not studied the extent to which the tools and procedures available for legislative evaluations are applied in practice by the DGs, nor the question for which aims legislative evaluations are used in the Commission. These topics would be worthy of further investigation.

Second, future research could take a look at the consequences of capacity differences for the initiation and the quality of legislative evaluations. From a technocratic perspective we could expect that DGs which invest more human and technological capital in legislative evaluations (i.e. DGs with high evaluation supply) produce more and better evaluations. From a more political perspective, we could expect that DGs which attach more value to evaluation (i.e. DGs with high evaluation demand) show better evaluation outputs (Mastenbroek et al., 2016). By further studying these topics, the data on evaluation capacity presented in this article could help to enhance our understanding of both the nature of the European Commission and the role of legislative evaluations in the European policy cycle.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. This concerned the DGs ESTAT (phone), ECFIN (phone) and MOVE (email).
2. Respondents from DG ENTR, MARKT and MOVE were unable to provide average evaluation costs, so for these DGs this information was collected by taking the average of three cost indications mentioned in tender or contract specifications published at http://ec.europa.eu/transport/facts-fundings/tenders/index_en.htm
3. Annual activity reports for 2013-2014 are available at http://ec.europa.eu/atwork/synthesis/aar-archived/aar_2013_en.htm. Activities categorized as both spending and non-spending activities were counted as half a spending activity.
4. The analysis was mostly conducted with the fuzzy add-on in Stata. For details about this add-on, see http://www.stata-journal.com/sjpdf.html?articlenum=st0140
5. Based on 2015 staff figures: http://ec.europa.eu/civil_service/docs/europa_sp2_bs_cat-sexe_x_dg_en.pdf
6. The intermediary solution, which is the recommended approach for fsQCA in most circumstances (Ragin, 2008, 164), only takes configurations which do not appear in the data into account if they meet certain assumptions. Following the theoretical framework of the article, positive relations were assumed between high amounts of legislation and high budgets and the outcomes, while negative relations were assumed between high political sensitivity and the outcomes.
7. The cut-off point for the amount of legislation was decreased and increased by up to 20 pieces of legislation; the cut-off point for budgets was decreased and increased by up to €500 million. In both cases, the most extreme cut-off points tested left only a few cases above or below their values.

References

Borrás S and Højlund S (2015) Evaluation and policy learning: The learners' perspective. *European Journal of Political Research* 54(1): 99–120.

Bourgeois I and Cousins JB (2013) Understanding dimensions of organizational evaluation capacity. *American Journal of Evaluation* 34(3): 299–319.

Bovens M, Hart P 't and Kuipers S (2008) The politics of policy evaluation. In: Goodin RE, Rein M and Moran M (eds) *The Oxford Handbook of Public Policy*. Oxford: Oxford University Press, 320–35.

Boyle R, Lemaire D and Rist RC (1999) Introduction: Building evaluation capacity. In: Boyle R and Lemaire D (eds) *Building Effective Evaluation Capacity: Lessons from Practice*. New Brunswick, NJ: Transaction Publishers, 1–19.

Carman JG and Fredericks KA (2010) Evaluation capacity and nonprofit organizations: Is the glass half-empty or half-full? *American Journal of Evaluation* 31(1): 84–104.

European Commission (2007) *Communication from the Commission from ms Grybauskaité in Agreement with the President. Responding to Strategic Needs: Reinforcing the Use of Evaluation. [SEC(2007)213]*. Brussels: European Commission.

European Commission (2013) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Strengthening the Foundations of Smart Regulation: Improving Evaluation. [COM(2013)686]*. Brussels: European Commission.

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

Fitzpatrick T (2012) Evaluating legislation: An alternative approach for evaluating EU internal market and services law. *Evaluation* 18(4): 477–99.

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

Kogut B, McDuffie JP and Ragin CC (2004) Prototypes and strategy: Assigning causal credit using fuzzy sets. *European Management Review* 1: 114–31.

Majone G (1999) The regulatory state and its legitimacy problems. *West European Politics* 22(1): 1–24.

Mastenbroek E, van Voorst S and Meuwese ACM (2016) Closing the regulatory cycle? A meta-evaluation of ex-post legislative evaluations by the European Commission. Epub ahead of print 5 October 2015. DOI: 10.1080/13501763.2015.1076874.

Nielsen SB, Lemire S and Skov M (2011) Measuring evaluation capacity: Results and implications of a Danish study. *American Journal of Evaluation* 32(3): 324–44.

Nugent N (2010) *The Government and Politics of the European Union*, 7th edn. Houndmills: Palgrave.

Pattyn V (2014) Why organizations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods. *Evaluation* 20(3): 348–67.

Preskill H and Boyle S (2008) A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation* 29(4): 443–59.

Ragin CC (2008) *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago, IL: Chicago University Press.

Ramboll Management Consulting (2011) *The Evaluation Capacity Index*. Available on request from EvaluationSociety@r-m.com.

Scharpf FW (1999) *Governing in Europe: Effective and Democratic?* Oxford: Oxford University Press.

Stern E (2009) Evaluation policy in the European Union and its institutions. In: Trochim WMK, Mark MM and Cooksy LJ (eds) *Evaluation Policy and Evaluation Practice: New Directions for Evaluation*. San Francisco, CA: Jossey-Bass, 67–85.

Stockdill SH, Baizerman M and Compton DW (2002) Toward a definition of the ECB process: A conversation with the ECB literature. *New Directions for Evaluation* 93: 7–25.

Summa H and Toulemonde J (2002) Evaluation in the European Union: Addressing complexity and ambiguity. In: Furubo J, Rist RC and Sandahl R (eds) *International Atlas of Evaluation*. New Brunswick, NJ: Transaction, 407–24.

Taut S (2007) Studying self-evaluation capacity building in a large international development organization. *American Journal of Evaluation* 28(1): 45–59.

Taylor-Ritzler T, Suarez-Balcazar Y, Garcia-Iriarte E, et al. (2013) Understanding and measuring evaluation capacity: A model and instrument validation study. *American Journal of Evaluation* 34(2): 190–206.

Technopolis (2005) *Study on the Use of Evaluation Results in the European Commission*. Brussels: European Commission.

Toulemonde J, Summa-Polit H and Usher N (2005) Triple check for top quality or triple burden? Assessing EU evaluations. In: Schwartz R and Mayne J (eds) *Quality Matters: Seeking Confidence in Evaluating, Auditing and Performance Reporting*. New Brunswick, NJ: Transaction, 66–90.

Weiss CH (1993) Where politics and evaluation research meet. *American Journal of Evaluation* 14(1): 93–106.

Stijn van Voorst is a PhD candidate at Tilburg University and Radboud University, the Netherlands. He works on a four-year project about ex-post legislative evaluation in the EU and has published in journals such as *European Journal of Risk Regulation*, *International Review of Administrative Sciences* and *Journal of European Public Policy*.